DOCUMENT RESUME

ED 267 349 CG 018 929

AUTHOR Goh, David S.
TITLE Applying Criterion-Referenced Measurement to School Psychology: A Handbook.
PUB DATE Apr 85
NOTE 100p.; Paper presented at the Annual Meeting of the National Association of School Psychologists (17th, Las Vegas, NV, April 8-12, 1985).
PUB TYPE Guides - Non-Classroom Use (055) -- Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Counseling Techniques; *Counselor Training; *Criterion Referenced Tests; Higher Education; Measurement Techniques; *School Counseling; *School Psychologists; Test Reliability; *Test Selection; *Test Validity

ABSTRACT

This document is a training module designed to familiarize practicing school psychologists and school psychology students with the field of criterion-referenced measurement and recent developments within the field. The introduction gives reasons for the need for criterion-referenced testing as opposed to norm-referenced tests which include legal requirements to gather data for prescriptive-remedial purposes. The first unit introduces the theory, background, and applications of criterion-referenced tests including the conceptual background, definitions, and the distinction between criterion-referenced and norm-referenced evaluation. The second unit discusses test development and psychometric aspects of criterion-referenced tests including test construction, reliability, validity, and standard setting. The third unit discusses evaluation and selection of criterion-referenced tests and the fourth unit contains descriptions and reviews of 20 popular, commerically available criterion-referenced tests in the field of school psychology. Six pages of references are included. (ABL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED267349

Applying Criterion-Referenced Measurement to School Psychology: A Handbook

David S. Goh

Southern Illinois University - Carbondale

CG018929

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

David S. Goh

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TABLE OF CONTENTS

INTRODUCTION..... 1

UNIT I. THEORY, BACKGROUND, AND APPLICATIONS..... 4

Conceptual Background..... 4

Definitions..... 5

Distinction Between Criterion-Referenced and Norm-Referenced Evaluation..... 8

Applications.....13

UNIT II TEST DEVELOPMENT AND PSYCHOMETRIC ASPECTS 18

Test Construction..... 19

Reliability..... 24

Validity..... 27

Standard Setting..... 31

UNIT III EVALUATION AND SELECTION OF CRITERION - REFERENCED TESTS 37

UNIT IV REVIEWS OF SELECTED CRITERION - REFERENCED TESTS..... 43

REFERENCES..... 84

FIGURE CAPTIONS..... 90

FIGURES 91

Introduction

The past decade has witnessed considerable controversy regarding the merits and limitations of the psychometric approach to assessment and diagnosis in the schools. The need for the school psychologist go beyond the traditional norm-referenced approach to gather data for prescriptive-remedial purposes has been repeatedly documented in the professional literature, including PL 94-142, the Education for All Handicapped Children Act, and the Larry P. decision. Consequently, criterion-referenced testing, among others, has been considered a possible alternative. Historically, criterion-referenced evaluation as a measurement model has been discussed rather extensively in the field of measurement and education, especially during the past ten to fifteen years. However, it has received far less attention in the field of school psychology. In a recent survey, Goh, Teslow, and Fuller (1981) have reported that a great majority of the assessment devices which are being used by school psychologists in the nation are norm-referenced tests; very few are criterion-referenced measures. Part of the reason, obviously, is due to the lack of well-established criterion-referenced instruments which are particularly pertinent to the practice of school psychology. Nevertheless, the lack of development of criterion-referenced assessment in the field of school psychology seems also attributable to the unfamiliarity of the school psychologist with the concepts and procedures of criterion-referenced measurement. In as much as criterion-referenced measurement has established itself as a test theory independent of the classical norm-referenced theory, a great deal of misconceptions exist among psychologists concerning the characteristics and proper uses of criterion-referenced evaluation. Many do not seem to realize that criterion-referenced and norm-referenced evaluations are quite different measurement

systems in their construction, interpretation, and evaluation. As a result, they tend to readily apply their knowledge of norm-referenced testing (i.e. reliability, validity, etc.) to criterion-referenced evaluation, assuming automatic applicability. Meanwhile, a great number of the commercially available criterion-referenced instruments are "home-made" without giving the due consideration to the required methodology of criterion-referenced test construction. These tests, as Popham (1978) has pointed out, are criterion-referenced tests in name only; their results are not useful and, sometimes, can be quite misleading.

It is our belief that criterion-referenced evaluation, when properly implemented, can produce meaningful contributions to the practice of school psychology (i.e. assessment, program evaluation, research, etc.). To this end, we have developed this training module to familiarize practicing school psychologists and school psychology students with the field of criterion-referenced measurement and its recent developments. Since most literature in this area has been written by measurement specialists in highly technical forms, efforts were made to present the measurement issues in nontechnical terms which are readily understandable by/and of practical value to the school psychologist as a practitioner. The purpose is to provide a systematic set of material which can be easily comprehended by the intended reader and help to prepare him/her with the proper concepts and procedures of criterion-referenced test development, selection, use, and evaluation. The training module includes four main units as follows:

I. An introduction to the theory and background of criterion-referenced evaluation and its potential applications in school psychology.

II. A review of the criterion-referenced measurement technology, in-

cluding test construction, reliability, validity, and other psychometric considerations.

III. A presentation of specific guidelines for the evaluation and selection of criterion-referenced devices.

IV. A description and review of selected criterion-referenced tests which are of interest to school psychologists.

Unit 1: Theory, Background, and Applications

Conceptual Background

The concept of criterion-referenced evaluation originated in the field of education. As early as 1918, E. L. Thorndike noted "there are two somewhat distinct groups of educational measurement: one. . . asks primarily how well a pupil performs a certain uniform task (norm-referenced); the other. . . asks primarily how hard a task a pupil can perform with substantial perfection, or with some other specified degree of success (criterion-referenced)" (Thorndike, 1918, p. 18). However, it was not until the 1960s that this approach of evaluation was formally developed and studied by measurement specialists. Glaser's article (1963) entitled "Instructional technology and the measurement of learning outcomes" was considered the first real harbinger of the criterion-referenced movement. In this article, Glaser introduced the expression criterion-referenced measurement and developed the distinction between criterion-referenced and norm-referenced types of tests, "What I shall call criterion-referenced measures depend upon an absolute standard of quality, while what I term norm-referenced measures depend upon a relative standard. . ." (p. 519). He indicated that knowledge acquisition can be measured on a continuum ranging from zero proficiency to perfect performance. Along this continuum, criterion levels can be established at any point in instruction, and individuals' achievement levels, as indicated by the behaviors they displayed during testing, can be evaluated and referenced accordingly. A central notion was thus established in criterion-referenced evaluation, that the standard against which an individual's performance is compared is the predetermined behavioral criterion, not the performance of other

individuals.

Definitions

Since Glaser's article, a larger number of definitions of criterion-referenced evaluation have appeared in the professional literature. Gray (1978) analyzed 57 descriptions of criterion-referenced evaluation and concluded that there are many definitions of criterion-referenced evaluation rather than a single, agreed-upon definition. During the first ten years or so, (approximately 1963-1975) of the criterion-referenced testing movement, there seemed to be considerable mis-interpretation and confusion about Glaser's statements concerning "standard of quality." It was not meant to be equating criterion with proficiency levels, which seemed to permeate in some of the definitions contributed by others. In their 1969 paper, Popham and Husek described criterion-referenced tests as "those which are used to ascertain an individual's status with respect to some criterion, i.e., performance standard" (p. 1). The term "performance standard" seemed to compound the confusion. Popham (1974) later recognized this misunderstanding and explained that performance standard refers to a class of behaviors and not a cut-off score.

Glaser and Nitko (1971), in their widely cited definition, also use the term "performance standards" and refer to it as a class or domain of tasks. "A criterion-referenced test is one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards. Performance standards are generally specified by defining a class or domain of tasks that should be performed by the individual" (p. 653).

There are also definitions which do not address the performance standard issue. Harris and Steart (1971) stated, "A pure criterion-referenced test is one consisting of a sample of production tasks drawn from a well defined population of performances, a sample that may be used to estimate the proportion of performances in that population at which the student can succeed" (p. 1). Iven's (1970) definition also did not mention performance standards and stated that a criterion-referenced test "comprised of items keyed to a set of behavioral objectives" (p. 2).

The most widely accepted definition was later offered by Popham (1974, 1975): "A criterion-referenced test is used to ascertain an individual's status with respect to a well defined behavior domain" (p. 130). What Popham refers to is a behavior domain as opposed to a criterion; this domain meaning a class of behaviors. (A criterion is only one particular behavior or objective, but a criterion-referenced test usually measures a number of behaviors or objectives.)

According to Glass (1978), it was up until 1976 that there was the confusion whether to equate criterion to a performance standard, a proficiency level, or a cut-off score. However, currently there is a general agreement that a criterion-referenced test is intended to reference an individual's score to a well defined domain of behaviors.

Linn (1982) recently noted the diversity of terminology used in the field of criterion-referenced testing. "Criterion-referenced, domain-referenced, objective-referenced, competency, and mastery have been used interchangeably by some people but not by others as qualifiers denoting the type of test" (p. 12). On a broad, conceptual level, these are

all "criterion-referenced" measures. Criterion-referenced tests, domain-referenced tests, objective-referenced tests, are the three main kinds of tests. Sanders and Murray (1976) offered the following definitions for these tests:

> Criterion-referenced testing (CRT). Performance on a task is interpreted against an absolute standard without referencing the performance of others. Example: The standard for good performance is getting at least 80 percent of the test items correct on the criterion-referenced test. Johnny got 85 percent of the items correct so we conclude he is performing acceptably.
>
> Objectives-referenced testing (ORT). Performance on a test is interpreted by referencing the behavioral objective for which it was written. Example: Johnny got 75 percent of the test items which were written for a particular objective correct.
>
> Domain-referenced testing (DRT). Performance on a task is interpreted by referencing a well defined set or class of tasks (a domain). Example: We have selected a random sample of 10 test items from a pool of 100 items for basic fifth grade spelling. Johnny spelled nine out of 10 words correctly so we estimate that if tested over and over again he would be able to spell 90 percent of the spelling words correctly. (p. 18)

According to Hambleton, Swaminathan, Algina and Coulson (1978), if one adopts Popham's (1975) definition of a criterion-referenced test, there is basically no difference between criterion-referenced tests and domain-referenced tests. In the case of objectives-referenced tests, there is no domain of behaviors specified, and items are not considered to be representative of any behavior domain. Therefore, the types of generalizations that can be made are much more limited on objective-referenced tests than on criterion-referenced tests. For this reason, Hambleton, et al. (1978) recommended the use of criterion-referenced tests, with Popham's definition.

Distinction Between Criterion-Referenced and Norm-Referenced Evaluation

The differences between criterion-referenced evaluation and norm-referenced evaluation have been substantially written by measurement specialists (i.e., Block, 1971; Ebel, 1971; Glaser and Nitko, 1971; Hambleton and Novick, 1973; Harris, et al., 1974; Messick, 1975; and Popham and Husek, 1969). The two types of measurement represent two ways of expressing the quantity of an individual's performance. Generally speaking, the conception of norm-referenced evaluation is based on the psychological model of individual differences and normal distribution. In this model, an individual's performance, whether a psychological trait or an achievement level, is measured on a relative basis. The focus, therefore, is to compare the individual's performance to that of others or some normative group. On the other hand, criterion-referenced evaluation is based on the educational notion of teaching-learning relationship. In this model, an individual's performance, mainly educational achievement,

is measured on an absolute basis. The focus of evaluation is on the determination of the quantity or the degree to which the individual has attained the predetermined criterion.

Insert Figure 1 about here

A careful examination of the various aspects of criterion-referenced measurement and norm-referenced measurement reveals that the differences between the two do not relate much to the nature, content, or format of the test; rather the differences are primarily in the use and interpretation of results derived from the tests. Popham and Husek (1969) distinguish between norm- and criterion-referenced measures in terms of how they are used:

> Norm-referenced measures are those which are used to ascertain an individual's performance in relationship to the performance of other individuals on the same measuring device. (p. 2)
>
> Criterion-referenced measures are those which are used to ascertain an individual's status with respect to some criterion, e.g., performance standard. (p. 2)

In a norm-referenced situation, the interpretation of an individual's score, as referred to earlier, is made possible by comparing it with that of other people (i.e., Johnny's score was better than 70 percent of other children his age in the standardization population). In a criterion-referenced situation, however, the meaning of an individual's

score is derived from its comparison with some established behavior criterion (i.e., Johnny scored 90 percent correct on the test). The difference in the interpretation of results is attributed to the different standards of comparison employed in the two measurement systems. Thus, if one is interested in knowing how well an individual can perform in an inter-individual differences sense, one would want to choose a norm-referenced measure. On the other hand, if one is mainly concerned with determining what or how much an individual has learned or attained in terms of specific behaviors or skills, (s)he would use criterion-referenced tests. Selection of tests and/or the appropriate score interpretation, therefore, depend upon the objective for testing.

Due to the different nature of criterion- and norm-referenced measures, the two types of tests also differ considerably in many technical aspects, including test construction, variability in scores, reliability, validity, etc. Recent advances in criterion-referenced measurement technology have formed a rather sophisticated test theory in its own right, independent of the classical norm-referenced test theory. A main difference underlying the two test theories is the issue of score variability. Score variability is central to the classical test theory. Norm-referenced tests are constructed in such a way that the range of scores that may be covered by a group of individuals is maximized. The idea is to show small gradations in differences among individuals. As Popham and Husek (1969) stated, test constructors "want variability and, as a consequence, make all sorts of concessions, sometimes subtle, sometimes obvious, to promote variant scores. He disdains items which

are 'too easy' or 'too hard.' He tries to increase the allure of wrong answer options" (p. 4). Variability in scores is not a critical consideration in criterion-referenced measurement, since the emphasis is not to discriminate between individuals but to determine whether or not an individual has achieved the established criterion. If all individuals achieved perfect scores on a criterion-referenced test, score variability will be reduced to zero. The degree of score variability does not affect test properties in a criterion-referenced situation, and no effort is made to promote or maximize score variability in such a situation. Instead, the central importance of a criterion-referenced test lies in content validity. That is, a sound criterion-referenced test should contain items that sample all of the important behaviors or objectives in the specified domain. In such instances, it is not appropriate to try to obtain some other items just because they may discriminate better. Item discrimination statistics should not determine the content of the test. These differences between criterion- and norm-referenced measurement directly affect many other technical aspects of the two types of tests. These issues will be discussed in a more extensive manner in Unit II of the training module.

Although measurement specialists generally agree on the above distinctions between criterion-referenced evaluation and norm-referenced evaluation, there appears to be some confusion and misunderstanding about the information each approach does or does not provide and the relationships between the two. For example, some seem to think that norm-referenced evaluation indicates nothing about what a person can do, but only about how (s)he compares with others, and that the opposite is true with criterion-referenced

evaluation. Many people tend to overlook the fact that criterion-referenced and norm-referenced evaluation are two different, but not necessarily mutually exclusive, measurement systems. In fact, the two approaches can best be used in a complementary manner. Ebel (1962, 1978) indicates that to be meaningful, any test scores must be related to the content of some specified domain (criterion-referenced) as well as the scores of other individuals (norm-referenced). Popham (1976) also suggests that normative data be considered in establishing performance standards in criterion-referenced tests.

> Suppose we have created a well defined criterion-referenced test of learners' attitudes toward school. If 100 points are the total possible when the child displays an oedipus-like attraction to the raptures of school, just how would we interpret a school district's average score of 62 points? Is 62 good or bad? Does it reflect a school perceived by students as Disneyland or a dung heap? Comparative data can help us answer such questions.
>
> If test performance was crisply described before the addition of norm data, then that crisp description won't wilt in the presence of norm data. In other words, you don't LOSE clarity of description by augmenting a test with comparative data, you merely pick up some information that's useful in setting reliable performance expectations. (p. 594)

Insert Figure 2 About Here

Applications

The nature of criterion-referenced evaluation suggests that it is most useful in describing an individual's status--what (s)he can or cannot do--with respect to specified behavioral objectives or in interpreting the individual's progress in terms of those objectives. To date, the primary application of criterion-referenced evaluation has been in several areas of education. Prominent among these are mastery testing, individualized instruction, and other similar instructional systems (computer-assisted, computer-managed, self-paced, etc.). In mastery testing, a person is determined either to have achieved (mastered) the objectives satisfactorily or to have not achieved. (A mastery test is a special type of criterion-referenced test.) Closely linked to mastery testing is the construction of individual teaching programs. In these programs, the instructional decisions for students are dependent on their performance on mastery tests. A student is required to achieve a set of specified objectives and then proceeds to the next set of objectives. Information obtained from the mastery tests reveals specific difficulties students may be experiencing, which is followed by particular instructional objectives. Most individualized instructional programs are composed of units or modules arranged in an hierarchical order; each involves the learning and measurement of one or more behavioral objectives. Examples of programs utilizing mastery testing include Individually Prescribed Instruction (IPI) developed by the University of Pittsburgh and PLAN (Planning for Learning in Accordance with Needs) by the American Institute for Research.

A great portion of the published criterion-referenced achievement tests are designed for group use, although individual use is also possible. These tests have been found useful in instructional decision making within the classroom for the purpose of determining student progress, curriculum development, and evaluation of instructional programs. When a decision regarding a student or group of students is required, a criterion-referenced test may be used to determine whether the student has acquired the learning skills considered to be prerequisites to more advanced skills or a new program. Such information obviously would be very helpful to the teacher in identifying missing skills as well as in planning appropriate programs for the student(s).

In order to evaluate instructional programs, it is necessary to have data about the outcomes on the specific objectives the program was designed to teach. A criterion-referenced measure which contains the program objectives to be achieved to those who have completed the program will give the necessary effectiveness data. Berguist and Graham (1980) described a program designed to evaluate a special education project using criterion-referenced tests. The evaluation included pre- and post-treatment effects. Comprehensive tests containing one or more criterion-referenced items for each objective were used as the pretests and posttests. Children identified and placed into a special education program were expected to meet a criterion of mastery of specified objectives during the period of evaluation. The degree of mastery of the objectives determined the effectiveness of the program.

Recently, criterion-referenced tests have drawn considerable attention from special educators as a result of PL 94-142, the Education for All

Handicapped Children Act. PL 94-142 mandates that an Individualized Educational Plan (IEP) must be implemented for each handicapped student placed in the special education programs. The IEP should specify, among other things, both instructional goals and objectives for the student, as well as criteria and procedures for evaluating the student's progress. Clearly, criterion-referenced evaluation, by nature, would serve this need more effectively than traditional norm-referenced evaluation.

Criterion-referenced evaluation can play a meaningful role in the practice of school psychology. It has the potential of making significant contributions to a number of important functions of the school psychologist, including assessment, consultation, program evaluation, and research.

Whereas currently popular norm-referenced tests can be used to identify students who are sufficiently different from their age peers to require special education placement, criterion-referenced tests will provide additional information in the diagnostic-prescriptive aspect for the school psychologist. Incorporating criterion-referenced evaluation into the routine assessment practice will add a useful dimension to the data-gathering process for the development of remedial and intervention purposes and should increase the accountability of the school psychologist in service delivery. In addition, criterion-referenced theory and methodology may also help the school psychologist develop alternative assessment strategies. For example, Wendt (1978) discusses the notion of criterion-cognitive developmental assessment. In this approach, he suggests the application of criterion-referenced methodology to the assessment of cognitive processing abilities such as Piagetian concepts (i.e., classification, seriation, conservation, etc.).

An important issue in diagnostic evaluation is bias in assessment and placement of children with special needs. Criterion-referenced evaluation is relevant to the concept of nonbiased assessment with minority or disadvantaged group children, in that it emphasizes the identification of special needs of children rather than the labeling process. Increased use of criterion-referenced measures in the so-called multifaceted nonbiased assessment practice is expected. Bailey & Harbin (1980) have indicated that criterion-referenced evaluation can be most useful in such situations when the following conditions are met: "1. The importance of the skills measured by the instrument and taught in the curriculum are agreed upon by culturally diverse groups within the school system. 2. Criterion referenced items are constructed so as not to measure the skills of children from a particular cultural group unfairly. 3. Alternative instructional strategies are incorporated to meet the learning needs of individual children." (p. 593)

Criterion-referenced evaluation also can contribute to the role of the psychologist as a program evaluator and/or a consultant in the schools. For example, school psychologists are often involved in the development and evaluation of intervention programs for a student or group of students. By carefully selecting a well designed criterion-referenced measure whose objectives match with those of a particular program, the school psychologist can appropriately determine the treatment outcome of the program. Finally, the role of the school psychologist as a consultant to both regular and special education programs has been repeatededly called for in the professional literature (Bardon and Bennet, 1974; Goh, 1977; Alpert, 1978). Criterion-referenced evaluation data should provide useful information

to the school psychologist in providing consultation about curriculum development and instructional planning and management within both regular and special education.

Unit II: Test Development and Psychometric Aspects

Since Popham & Husek's (1969) now classical article, considerable advancement has been made in the methodological aspect of criterion-referenced measurement. Through the efforts of many measurement specialists, a fairly advanced theory of criterion-referenced measurement technology has gradually been developed. In the meantime, numerous criterion-referenced tests, almost exclusively in the field of achievement testing, have been published and used in the nation's schools. However, professional educators and psychologists do not seem to be highly aware of the technical developments of criterion-referenced measurement, although they are often involved in the development and use of criterion-referenced tests. In fact, a great number of the commercially available criterion-referenced tests reviewed in Buro's Mental Measurement Year Book and elsewhere (i.e., journals, etc.) were found to be of low quality, mainly because they were constructed without giving due consideration to the required methodology involved in criterion-referenced test construction. As a result, these tests are criterion-referenced tests in name only and their usefulness is quite limited (Popham, 1978).

School psychologists, perhaps, are the professionals most knowledgeable on measurement matters in a school or mental health setting. Other educational and health service providers often seek consultation from the school psychologist on issues relating to selection and uses of measurement devices for different purposes. As criterion-referenced tests are being used in increasing numbers in the schools, it is essential that the school psychologist become familiar with the measurement model underlying these

instruments. As noted, currently available criterion-referenced tests focus almost exclusively on the measurement of educational subject matters. However, it is believed that the concept of criterion-referenced evaluation may also apply to the measurement of psychological or behavioral constructs (i.e., adaptive behavior, social skills, etc.) and that more research is needed in the development of such instruments. This may prove an area in which the school psychologist can make significant contributions.

Classical test theory, which has been well developed for norm-referenced measurement, is not appropriate for criterion-referenced evaluation. The issue of variability is a central difference between the two approaches (Popham & Husek, 1969). While norm-referenced evaluation attempts to maximize score variability, no such effort is made in criterion-referenced measurement. The main purpose of criterion-referenced measurement is not to discriminate between individuals as is done with norm-referenced instruments, but rather its major purpose is to describe an individual's performance with respect to well defined objectives or criteria. This shift of measurement focus affects many technical aspects of criterion-referenced measurement and, thus, calls for a test theory of its own.

Test Construction

A basic principle in any test construction is that a test should be constructed in a manner appropriate to its intended purpose and use. In the development of a criterion-referenced test, it is essential to define operationally the domain of content or behaviors the test is to measure. It is also crucial that all test items be carefully made to represent

the domain of content or behaviors delimited by the criterion so that accurate inferences can be made from the test results. The validity and interpretability of test results are determined by the precision of the behaviors defined and domain specified. Hambleton (1980) suggests the following steps in the development and validation of criterion-referenced tests.

1. Objectives or domain specifications must be prepared or selected before the test development process can begin.
2. Test specifications are needed to clarify the test's purposes, desirable test item formats, number of test items, instructions to item writers, etc.
3. Test items are written to measure the objectives included in a test (or tests, if parallel forms are required).
4. Initial editing of test items is completed by the individual s writing them.
5. A systematic assessment of items prepared in steps 2 and 3 is conducted to determine their match to the objectives they were written to measure and to determine their "representativeness."
6. Based on the data from step 5, additional item editing is done. Also, test items are discarded that do not at least adequately measure the objectives they were written to measure.
7. The test(s) is assembled.

8. A method for setting standards to interpret examinee performance is selected and implemented.
9. The test(s) is administered.
10. Data addressing reliability, validity, and norms are collected and analyzed.
11. A user's manual and technical manual are prepared.
12. A final step is included to reinforce the point that it is necessary, in an ongoing way, to be compiling technical data on the test items and tests as they are used in different situations with different examinee populations (from Berk, 1980, pp. 81-82).

A well developed criterion-referenced test must include these steps in its construction, although not necessarily in the exact order. Domain specification and test specification are the most important considerations in the initial stage of test construction. Domain specification refers to specially prepared statements which describe in nonambiguous terms the content implied by the domain. (A domain or behavior objective is a set of well defined skills or class of behaviors to be measured by the test.) Test specification consists of a set of rules which are needed to create test items that would be representative of the behaviors or skills identified by the behavior domain(s). The terms of behavioral objectives, domain statements, domain specifications, and test specifications sometimes are used interchangeably by different writers. A number of domain specification strategies have been developed over the past decade, including such notions as behavioral objectives and amplified objectives developed

by Popham and his associates (i.e., Popham, 1974), item-form analysis by Hively and his associates (Hively, Patterson, and Page, 1968), and facet analysis by Berk (1978). Popham (1978, 1980) recently proposed a limited-focus strategy for developing test specifications which include four essential components. First, general description, also known as objectives in some tests. This refers to a brief general description of what it is that the test measures. Second, a sample item, complete with directions to the examinee. This will help to clarify the item domain specified by general description and serves to provide format cues for test item writers. Third, stimulus attributes; this section of test specifications contains a series of statements that attempt to delimit the content of stimulus material that will be encountered by the examinee. In other words, these statements specify the major rules for generating the content items which constitute the test. Finally, response attributes, a series of statements that attempt to specify response formats required of the examinee, guidelines for creating response alternatives, and criteria for judging the correctness of the examinee's responses. An example of these test specifications is provided in Figure 3. It should be noted that none of the domain specification or test specification strategies available in the literature are considered well established. They are basically conceptual recommendations by measurement specialists which await empirical validation. Domain specification is a crucial step in criterion-referenced test construction and a difficult task for the test constructor. What is most important to remember is that a domain must be well defined in objective, behavioral terms and that specific rules need to be carefully established regarding the indispensable elements that item writers must consider in producing test items (Popham, 1978).

Only based on clearly stated domain and test specifications can functional test items be generated to accurately reflect the behavioral domains to be measured by the test.

Insert Figure 3 about here

Based on the prepared domain and test specifications, a large pool of homogeneous items can be developed. Normally, an item pool is generated for each single behavioral objective or domain identified. Two requirements should be observed in item development and item selection for the test: a) Every item that could be written from the domain must be written (or known) before item selection, and b) The criterion exercise must constitute a random sample from the item population (Ebel, 1962; Hively, Patterson, and Page, 1968). Many test authors used a rather subjective procedure (i.e., own or expert judgment) for selecting items to be included in a test. However, a random procedure is more appropriate. There are two different types of random procedures which can be used to select test items from a large item pool. These are simple random sampling and stratified-random sampling. When there is only a relatively small single domain and a fairly large number of items have been created, the simple random sampling method is appropriate. However, if there are large and somewhat related domains, then stratified-random sampling will be more desirable in generating representative samples of test items. In stratified-random sampling, the item pool is subdivided in different strata from which items are randomly sampled. These strata are typically formed according to item content, item-difficulty level, or objective. Shoemaker (1972) states, "Because the item population in criterion-referenced measurement is usually

not specified completely and random sampling is neglected (in favor of expert selection), the proportion score is frequently meaningless." (p. 38) He further suggests that items should be stratified and selected in such a way that, "(a) a proportion of the items will be answered correctly by all examinees achieving the minimal level of satisfactory performance, (b) a proportion of the items will be answered correctly only by those examinees who have surpassed the minimal level of achievement, and (c) the remaining items will be answered correctly only by those examinees achieving a high level of mastery on that objective." (p. 61-62)

In addition to domain specification and item generation and selection, the test constructor needs to be concerned with a number of things in test assembly. These include determining test length, preparation of directions, scoring keys, etc. (Hambleton, 1978). The issue of test length has been discussed rather extensively in criterion-referenced measurement (cf. Hambleton, et al., 1978). Obviously, neither too many nor too few items are desirable. While there is no simple answer to this issue, as a rule, Popham (1978) has suggested that 5 to 20 items per measured behavior seem adequate.

Reliability

Reliability in norm-referenced tests refers to the consistency of test scores and is usually estimated by correlational procedures (i.e., KR-20, etc.). These procedures, however, are not well suited for use with criterion-referenced tests, mainly because of the reduced variance issue in these tests. While a certain amount of response variance is expected on any type of test, a criterion-referenced test makes no attempt to maximize score variability. Therefore, the use of the traditional methods for estimating reliability would produce spurious correlations which are

difficult to interpret. Popham (1978) indicated that the correlational methods are appropriate only if a significant amount of score variability is evident on a criterion-referenced test. Most measurement experts seek out other methods for estimating reliability of criterion-referenced tests.

Reliability in criterion-referenced tests deals with the consistency of the decisions resulting from test scores, not merely the test scores themselves. On each criterion behavior measured by a criterion-referenced test, a cut-off score (i.e., median, 85 percent correctness, etc.) can be established to classify individuals into two or more gross categories. By comparing an examinee's score with the cut-off score, a decision can be made regarding the examinee's performance status in relationship to the cut-off score (i.e., pass-fail, mastery, partial mastery, non-mastery). Reliability is then determined by examining how consistently the test classified examinees into the same category on separate occasions (test-retest). Likewise, reliability can be determined for a test using equivalent forms. Three rather simple and easy to use methods for establishing reliability of the decision process are described here. In the percentage of decision-consistency method, the percentage of agreement of the decisions made regarding examinees' status (i.e., mastery, non-mastery) on two or more administrations of the same test is computed. The higher the percentage of agreement (the proportion of examinees classified in the same category), the higher the reliability of the test. This method is illustrated as follows:

		Administration 1	
		Masters	Non-Masters
Administration 2	Masters	Pa1	Pd2
	Non-Masters	Pd1	Pa2

Where Pa1 = Proportion of examinees classified as masters on both administrations.

Pa2 = Proportion of examinees classified as non-masters on both administrations.

Pd1 = Proportion of examinees classified as masters on first administration, but non-masters on second administration.

Pd2 = Proportion of examinees classified as non-masters on first administration, but masters on second administration.

Percentage of decision-consistency (P_o) equals the sum of Pa1 and Pa2. However, this value is most likely to represent an overestimate of the reliability of the decision results since it involves some amount of chance agreement. Swaminathan, Hambleton, and Algina (1974) suggested the use of coefficient K (Cohen, 1960), which is defined as follows, to control for chance agreements.

$$K = \frac{Po - Pc}{1 - Pc}$$

Where Po = Pa1 + Pa2

Pc (proportion of chance agreements)

= (Pa1 + Pd2) (Pa1 + Pd1) + (Pd1 + Pa2) (Pd2 + Pa2)

The second method employs non-parametric procedures to analyze the decision results (i.e., as shown in the aforementioned figure) and estimate reliability coefficients (Popham, 1978). For example, a chi-square test or a phi coefficient can be computed to estimate the degree of consistency to which the decisions are made based on the examinee's performance on different occasions or on equivalent forms of a test. The third method is to simply compare each examinee's score from the first test administration with his/her score from the second administration and determine how much they differ. The degree of reliability of the test can then be shown by reporting the percentage of examinees whose two scores differ by little (0 - 5%), slightly more (6 - 10%), and so forth. In addition to the methods described, more complex statistical analyses (Hambleton, 1978; Millman, 1979) have also been proposed for estimating reliability of a criterion-referenced test. However, Popham (1978) indicated methods like percentage of decision consistency are easy to use and can prove quite sufficient for most situations.

Validity

Validity refers to the degree to which a test measures what it purports to measure. Validity is one of the most important psychometric properties to any test. As commonly known, reliability is a prerequisite of validity; but high reliability does not necessarily guarantee high validity. A test may have high reliability, but without validity it is still useless. The validity issue in criterion-referenced measurement can be examined from several perspectives. The most essential type of validity is content validity, or descriptive validity, in Popham's (1978) terms. Content validity refers to the congruence between test content

and the behavior domain(s) defined in a test. For a sound criterion-referenced test, it is crucial that there are close relationships between the test items and the behavior domain(s) that the test was designed to measure. Hambleton, Algina, and Coulson (1978) indicate that content validity can be examined from a theoretical perspective in terms of test construction. That is, if a criterion-referenced test is developed strictly following the legitimate and necessary steps in test construction, content validity should not be a problem. Other writers also have pointed out the importance of establishing direct relationships between test items and domain(s) of interest.

Besides the theoretical framework, one way to empirically estimate content validity of a criterion-referenced test is through the use of content specialist ratings. This method involves having two or more content specialists judge the relevance of each item to the domain it is intended to measure. A value of +1 is assigned if the judge feels the item is definitely a measure of the behavior domain; a value of -1 is assigned if the item is judged not a measure of the domain; and a value of 0 is assigned otherwise (i.e., undecided). The resulting data then are used to compute an Index of Item - Objective (domain) Congruence (Rovinelli and Hambleton, 1977) which was derived from Hemphill-Westie Index (1950). The formula is as follows:

$$I_{io} = \frac{(M - 1)\ (S_o - S_{o'})}{2N\ (M - 1)}$$

Where I_{io} - the Index of congruence for item i and domain o.

M = the number of domains.

N = the number of specialists.

S_o = the sum of the item ratings assigned to domain o.

S_o1 = the sum of the ratings assigned to all domains except domain o.

The value of I_{io} ranges from -1.00 to +1.00. The higher the value of I_{io}, the higher the degree of congruence for each item-domain combination (item content validity). The degree of content validity of the whole test can then be estimated by determining the proportion of test items that is consistent with the domain description.

Although content validity is extremely important to criterion-referenced tests, it has only limited focus and value from a measurement point of view. Messick (1975) indicates that content validity, as defined in criterion-referenced evaluation, is basically a test characteristic and focuses upon test forms rather than test results, upon instruments rather than measurements. He points out inferences in educational and psychological measurement are made from test scores which are a function of examinee responses. Therefore, other types of validity based on the inspection of examinee performance need to be established empirically for a criterion-referenced test. Hambleton, Algina, and Coulson (1978) also observe: "Content validity issues are essential at the test development stage, and content validity of criterion-referenced test will influence the kind of test score interpretations that are possible. But it is most important to conduct construct validation studies for the intended interpretations of a set of scores. Construct validation studies will relate to the matter of 'meaning of scores'" (p. 39). Construct validity, also known as domain-selection validity in criterion-referenced measurement, is important as it indicates the degree to which the domain selected on a test will reflect examinee status with respect to the more general dimension

under consideration. In other words, it deals with the important question of generalizability of criterion-referenced test results. Only on a test with high construct or domain-selection validity can generalizations be properly made based on the examinee's status on the test. Empirical methods such as factor analysis as well as experimental procedures typically have been suggested for studying the construct validity of criterion-referenced test scores (i.e., Popham, 1978). For example, an experiment can be designed in which subjects are randomly assigned to two groups. One group receives instruction on the general content defined by a domain specification and the other does not. If treatment is effective, higher test scores by the experimental group would support the construct validity hypothesis (Hambleton, Algina, and Coulson, 1978). Likewise, a pre- and post-measure type of design can be used. Significant difference on the test scores in favor of the post-measure would be indicative of construct validity.

Recently, Swezey (1981) discusses two other types of validity as also important to a criterion-referenced test--concurrent validity and predictive validity. The concepts underlying these validities are the same as those in norm-referenced measurement; however, slightly different procedures are used for estimating these validities in a criterion-referenced situation. Both concurrent validity and predictive validity are criterion-related validity in that they reflect the relationships between a criterion-referenced test score(s) and scores on an external measure. The only difference between the two is on the timing of the administrations of the test and the external measure. In concurrent validity, both are administered at the same point in time; whereas in predictive validity, external data are always collected at some future time. Swezey (1981)

suggests the use of the statistic Ø to estimate concurrent validity as well as predictive validity. "If the Ø coefficient for concurrent (or predictive) validity is +.50 or above, the criterion-referenced test is of suitable validity. If the Ø coefficient is in the range between -1.00 and +.50, the test is of questionable validity" (p. 154).

Standard Setting

A major use of criterion-referenced tests is to assist in decision making. That is, to determine whether an individual has or has not achieved the performance standard within a behavior domain(s). This use is most common in pupil diagnosis with regard to predetermined learning objectives and in pupil certification with regard to comprehensive decision making such as grade-to-grade promotion or school graduation (Shepard, 1980). Decision making in criterion-referenced measurement involves the notion of setting performance standards or passing scores on a test. As mentioned earlier in this training module, criterion-referenced evaluation does not compare individuals on a relative basis. Rather, the meaning of criterion-referenced test results is derived on an absolute and axiomatic basis, by comparing an individual's performance with certain predetermined performance standard(s) or behavior criterion. Performance standard is central to the decision-making practice in criterion-referenced measurement and is generally considered as the distinguishing characteristic of criterion-referenced tests.

Despite its understood importance, setting performance standard or cutoff score has been a controversial issue in criterion-referenced evaluation. The main reason is that standard setting is judgmental and arbitrary. Glass (1978) started this controversy when he reviewed six

different methods for establishing performance standards or cutoff scores and concluded that all are arbitrary and insufficiently served the purpose of decision making. Consequently, he advocates that standard setting should be avoided. He states:

> I am confident that the only sensible interpretations of data from assessment programs will be based solely on whether the rate of performance goes up or down. Interpretations and decisions based on absolute levels of performance on exercise will be largely meaningless, since these absolute levels vary unaccountably with exercise content and difficulty, since judges will disagree wildly on the question of what consequences ought to ensue from the same absolute level of performance on exercises to succeed on the job, in higher levels of schooling, or in life. Setting performance standards on tests and exercises by known methods is a waste of time or worse. (p. 259)

Other measurement specialists (i.e., Block, 1978, Hambleton, 1978, Popham, 1978) acknowledge that standard setting is judgmental; however, they argue it is inevitable and perhaps an arbitrary standard is better than none at all. Ebel (1978) offers the following comments:

> Pass-fail decisions on a person's achievement in learning trouble some measurement specialists a great deal. They know about errors of measurement. They know that some who barely pass do so only with the help of errors of measurement: They know that some who fail do so only with the hindrance of errors of measurement. For these, passing or failing does not depend on achievement

> at all. It depends on luck. That seems unfair, and indeed it is. But, as any measurement specialist can explain, it is also entirely unavoidable. Make a better test and we reduce the number who will be passed or failed by error. But the number can never be reduced to zero. (p. 549)

Two different models and numerous methods for standard setting have been proposed. Hambleton, et al. (1978), Meskauskas (1976), Millman (1973), and Shepard (1979) all have provided excellent reviews of these methods. However, none of these proposed methods has been widely accepted. It is generally recognized that each method has its merits and limitations. Zieky and Livingston (1977) and Popham (1978) have produced some guidelines for the use of several more popular methods. But there is a general lack of effective guidelines for the selection of standard setting methods for particular situations. Shepard (1980) recently suggests that for some purposes, it is recommendable to use more than one method to compensate for the limitations in each method.

Standard setting methods are generally developed within two different theoretical frameworks. The first is known as the continuum model. In this model, it is assumed that the ability or behavior being measured is continuous in nature. Mastery based on this model is defined in a dichotomous decision by using a discrete cutoff score set by the test constructor or test user. The second model is the state model. In this model, mastery is defined by an all-or-none state. That is, an individual either has mastered the complete behavior or skill or he has not. Shepard (1980) indicated the state model has rather limited value, except for very tiny behavior domains such as single-digit

addition problems. The continuum model is more plausible and most standard setting methods proposed are based on this model.

Hambleton (1980) has classified methods for setting standards into three categories: judgmental, empirical, and combination methods. All

---

Insert Figure 4 about here

---

judgmental methods employ a common-sense approach. Judges who are highly knowledgeable about the behavior domain(s) of interest are asked to inspect test items and determine the minimal competence level for each item. Results provided by the judges are then pooled to obtain an estimation of the level at which the examinees should be expected to perform, such as percentage of correct answers needed for passing, etc. Setting performance standard in judgmental methods relies mainly upon test content and the opinions of the judges; it is not concerned with actual performances of examinees. It is critical that a number of judges should be used and the judges selected should be representative of the personnel usually involved in the particular decision making situation. Furthermore, there should be a high degree of agreement among the judges. Of the judgmental methods listed in Figure 4, the Nedelsky method was most often used in the health professions.

The empirical methods attempt to set performance standard based on actual data collected from examinees' performances on the test. Typically, an independent criterion measure is used to divide examinees into two groups (i.e., "instructed" vs. "uninstructed," etc.). The test is then administered to both groups and the cutoff score maximally separating

the groups is selected. These methods are based on decision theory and often employ mathematical models in determining performance standards. The main problem of these methods is that it would be extremely difficult to equate mastery with instruction and complete nonmastery with non-instruction, and, therefore, that the optimal cutoff score selected is really not a "true" standard (Shepard, 1980).

The combination methods select performance standard based on a combination of judgmental and empirical data. Judges are used first to determine the minimum performance level(s) for a particular behavior domain(s). They are then asked to identify individuals who show varying degrees of performance mastery (i.e., borderline groups, contrasting groups, etc.) of the behaviors measured. Empirical data are then collected for these groups and statistic procedures (Berk, 1976) are used to establish performance standards. The combination methods differ from the judgmental methods in that judgments are made not only about test content but also about individuals. In addition, technical data are gathered to improve standard setting in order to reduce false positives (assigning a nonmaster to a mastery state) as well as false negatives (assigning a mastery to a nonmastery state). Currently, most measurement specialists appear to favor the combination methods (i.e., Zieky & Livingston's Contrasting Groups method) in setting performance standards.

The issue of standard setting directly affects the practice of reporting test results. Obviously, notions such as percentile ranks and various standard scores which are widely used in reporting norm-referenced test results are inappropriate for criterion-referenced tests. Two types of scores are commonly used in a criterion-referenced situation:

"Level of functioning" (i.e., "pass-fail," "mastery, partial-mastery, non-mastery") and "percentage-correct" (i.e., proportion of test items answered correctly) scores. Both types of scores are determined by the performance standards set for a test. These scores should be reported for each of the behavior domains measured by a test, as well as for the whole test. Recently, Hambleton, Power, and Eignor (1979), Jaeger (1978), and Popham (1978) have suggested the use of norms in standard setting as well as reporting test results. Normative data concerning a group (or groups) of individuals' performances on each behavioral objective or domain can provide useful information for establishing sensitive and realistic performance expectations. Popham (1978) encourages test constructors to assemble normative data on various well described and defined groups for criterion-referenced tests. Furthermore, it should be noted that while numerical results or classification levels are useful for decision-making purposes, such results alone are insufficient in criterion-referenced evaluation. They should be accompanied by explicit statements describing the specific behaviors the individual can or cannot perform. A narrative report is most appropriate for presenting results from a criterion-referenced test. Preferably, target behavioral objectives, as well as a range of acceptance performance, should also be included in the report.

Unit 3. Evaluation and Selection of Criterion-Referenced Tests

A large number of criterion-referenced tests have been published since the early 1970s. As noted in Unit 2, considerable advances in criterion-referenced test technology were observed mainly during the past several years. Specific knowledge about criterion-referenced test construction and guidelines for test evaluation were initially lacking. As can be expected, the quality of the currently available criterion-referenced tests varies a great deal. These tests differ significantly from the degree to which they adequately sample the behavior domain(s) to the extent to which they meet the many essential technical requirements in test construction. It was evident many test authors and publishing companies, especially in the earlier days of the criterion-referenced testing movement, were not highly concerned about methodological considerations in developing their tests. As a result, the usefulness of these tests is very limited and the meaning of the results from these tests is often unclear, if not misleading. As such, it is only appropriate that test users be very careful in their evaluation and selection of criterion-referenced tests. Only those which are properly constructed to accomplish their intended purpose(s) should be selected for actual use. The Standards for Educational and Psychological Tests (APA, 1974) provides some general recommendations which should apply to norm-referenced as well as criterion-referenced tests; however, it does not contain guidelines specifically designed for the evaluation and employment of criterion-referenced tests. More recently, Hambleton and Eignor (1978), Popham (1978), Swezez and Pearlstein (1975), and Walker (1977) have proposed particular guidelines for evaluating criterion-referenced tests and test manuals.

Two sets of these guidelines are recommended in this module. Both should be useful to test users, as well as test constructors of criterion-referenced measures.

Popham (1978, pp. 177-184) discusses six important characteristics of a well constructed criterion-referenced test. These are:

1. An unambiguous descriptive scheme. This means the test should clearly state the procedure used to describe an examinee's performance on the test. A good criterion-referenced test should permit one to make an unequivocal description of what an examinee's performance truly means.
2. An adequate number of items per measured behavior. The optimum number of items per measured behavior may range from 5 to 20. Too many as well as too few items (i.e., less than 3 items per measured behavior) are inadequate.
3. Sufficiently limited focus. This refers to the number of behavior objectives measured by a test. Probably about 5 to 10 measured behaviors per subject per year is reasonable.
4. Reliability. A good criterion-referenced test should be highly reliable. That is, it should measure the behavior with considerable consistency.
5. Validity. The test should include evidence substantiating adequate descriptive, functional, and domain-selection validity.
6. Comparative test data. The test should be accompanied by field trial data indicating how other examinees performed on the test. Comparative data by geographic region, sex, age, etc., are highly recommended.

Hambleton and Eignor (1978) have developed a list of fairly elaborate guidelines for evaluating 10 different aspects of criterion-referenced tests. These are stated as follows:

A. Objectives

A.1 Is the purpose (or purposes) of the test stated in a clear and concise fashion?

A.2 Is each objective clearly written so that it is possible to identify an "item pool"?

A.3 Is it clear from the list of objectives what the test measures?

A.4 Is an appropriate rationale offered for including each objective in the test?

A.5 Can a potential user "tailor" the test to meet local needs by determining which objectives from a pool of objectives offered by the publisher are to be measured by the test?

A.6 Is there a match between the content measured by the test and situation where the test is to be used?

A.7 Are individuals identified who were responsible for the preparation of objectives?

A.8 Does the set of objectives measured by the test serve as a representative set from some content domain interest?

B. Test Items

B.1 Is the item review process described?

B.2 Are the test items valid indicators of the objectives they were developed to measure?

B.3 Is the set of test items measuring an objective representative of the "pool" of items measuring objective?

B.4 Are the items free of technical flaws?

B.5 Are the test items in an appropriate format to measure the objectives they were developed to measure?

B.6 Are the test items free of bias (for example, sex, ethnic, or racial)?

B.7 Was a heterogeneous sample of examinees employed in piloting the test items?

B.8 Was the item analysis data used only to detect "flawed" items?

C. Administration

C.1 Do the test directions include information relative to test purpose, time limits, practice questions, answer sheets, and scoring?

C.2 Are the test directions clear?

C.3 Is the test easy to score?

C.4 Does the test manual specify an examiner's role and responsibilities?

D. Test Layout

D.1 Is the layout of the test booklets attractive?

D.2 Is the layout of the test booklets convenient for examinees?

E. Reliability

E.1 Is the type of reliability information offered in the test manual appropriate for the intended use (or uses) of the scores?

E.2 Was the sample (or samples) of examinees used in the reliability study adequate in size, and representative of the population for whom the test is intended?

E.3 Are test lengths suitable to produce test with desirable levels of test score reliability?

E.4 Is reliability information offered in the test manual for each intended use (or uses) of the test scores?

F. Cut-Off Scores

F.1 Was a rationale offered for the selection of a method for determining cut-off scores?

F.2 Was the procedure for implementing the method explained, and was it appropriate?

F.3 Was evidence for the validity of the chosen cut-off score (or cut-off scores) offered?

G. Validity

G.1 Does the validity evidence offered in the test manual address adequately the intended use (or uses) of scores obtained from the test?

G.2 Is an appropriate discussion of factors affecting the validity of test scores offered in the test manual?

H. Norms

H.1 Are the norms data reported in an appropriate form?

H.2 Are the samples of examinees utilized in the norming study described?

H.3 Are appropriate cautions introduced for proper test score interpretations?

I. Reporting of Test Score Information

I.1 Are the test scores reported for examinees on an objective by objective basis?

I.2 Are there mutliple options available to the user for reporting of test results (for example, by class and grade within a school)?

I.3 Are convenient procedures available for scoring tests by hand, and forms available for reporting test score information?

J. Test Score Interpretations

J.1 Are suitable cautions included in the manual for interpreting individual and group objective score information?

J.2 Are appropriate guidelines offered in the manual for utilizing test scores to make descriptive statements, instructional decisions, program evaluation decisions, or other stated uses of the test scores? (pp. 322-324)

Unit 4 Review of Selected Criterion-Referenced Tests

This unit contains descriptions and reviews of the following more popular, commercially available criterion-referenced tests. The selection of these tests was based on a review of material included in Anastasi (1976), Buros (1978), Goh, Fuller & Teslow (1981), Hambelton & Eignor (1978), Salvia & Ysseldyke (1978) and professional journals in the fields of school psychology & education.

1. Classroom Reading Inventory
2. Criterion Test of Basic Skills
3. Diagnostic Mathematics Inventory
4. Diagnosis: An Instructional Aid--Mathematics
5. Diagnosis: An Instructional Aid--Reading
6. Fountain Valley Teacher Support System in Reading
7. Fountain Valley Teacher Support System in Mathematics
8. Individualized Criterion-Referenced Testing: Math
9. Individualized Criterion-Referenced Testing: Reading
10. Individual Pupil Mointoring System-Mathematics
11. Individual Pupil Monitoring System--Reading
12. Key Math Diagnostic Arithmetic Test
13. Mastery: An Evaluation Tool Mathematics
14. Mastery: An Evaluation Tool--Reading (SOBAR)
15. Prescriptive Reading Inventory

16. Wisconsin Design for Reading Skill Development: Word Attack
17. Wisconsin Design for Reading Skill Development: Study Skills
18. Woodcock Reading Mastery Tests
19. Standard Diagnostic Mathematics Test
20. Stanford Diagnostic Reading Test

Classroom Reading Inventory

A. General Information

1. Author: Silvaroli, N. J.

2. Publisher: Wm. C. Brown Company
135 South Locust St.
Dubuque, Iowa 52001

3. Cost: $1.95 per manual. The publisher gives permission to reproduce the Inventory Record, p. 17-26, so cost is minimal.

4. Date of Publication: 1976

B. Purpose and Nature of the Test:

The test is designed for use with students in grades 2-8. It is to be used to assess a students reading level and can provide information concerning word recognition and comprehension.

C. Development of the Test:

No information is provided in the manual concerning test development, standardization or norms.

D. Administration and Scoring:

The test can be individually administered in approximately 12 minutes, and no formal training is required. Some of the scoring procedures in the manual are unclear.

E. Analysis and Interpretation fo Results:

Six scores are obtained: word recognition, independent reading level, instructional reading level, frustration level, hearing capacity level, and spelling. Scores are transfered to a Student Inventory Record for interpretation.

F. Evaluation:

There is no reliability or validity information reported in the manual. Interpretations in a diagnostic sense should be made with caution. This test probably best serves as a rough screening device, and is quickly and easily administered.

Criterion Test of Basic Skills

A. General Information

1. Author(s): Lundell, K., Brown, W., and Evans, J.

2. Publisher: Academic Therapy Publications
1539 Fourth Street
San Rafael, California 94901

3. Cost: The testing materials are $17.00 per set. The set includes the manual and materials for 25 students and includes both the reading and math tests.

4. Date of Publication: 1976

B. Purpose and Nature of the Test:

The test was designed for use as a quick and easy method of assessing the basic reading and arithmetic skills of children in grades K-8. There is a total of 19 reading objectives dealing with letters (recognition, sounding, and writing), blending and sequencing, special sound, and sight words. The average number of items per objective is 13. The 26 math objectives deal with numbers and numerals, the four basic operations, and other application skills. The average item number per objective is about 6. Item formats include both verbal and performance responses.

C. Development of the Test:

The test is a criterion-referenced test with no field test data available. No information is provided to support the increasing difficulty level of items in the test, nor is there any information provided on the relevancy of the items tested to the grade levels specified.

D. Administration and Scoring:

The test is administered individually using a 4-page record sheet for each test. There are two separate tests--one math test and one reading test. Ten to fifteen minutes are required to administer each test. No special qualifications for the test examiner are mentioned. Scoring is done on the assessment records by determining the number of correct items in each section of the test.

E. Analysis and Interpretation of Test Results:

The reading test consists of 19 specific objective subtests in 6 areas. The arithmetic test consists of 26 specific objectives in 11 areas. Students take subtests beginning with 2 mastery scores and ending with 2 frustration scores. Mastery is defined as 90-100% correct, instructional level is defined as 50-89% correct, and frustration level is defined as 0-49% correct.

F. Evaluation:

No data was available on reliability, validity, or relevancy to grade level. The test can be administered by a classroom teacher within a 30-minute time period on an individual basis. The specific-objective subtests might help the classroom teacher identify skill weaknesses for certain individual students. The manual includes a comprehensive section of suggested teaching strategies.

Diagnostic Mathematics Inventory

A. General Information:

1. Author: Gessel, J. K.

2. Publisher: CTB/McGraw-Hill
Del Monte Research Park
Monterey, California 93940

3. Cost: The Diagnostic Mathematics Inventory comes in seven levels. Tests may be ordered in groups of 35 ranging from a cost of $19.25 for Level A, $20.30 for Level B, $23.80 for Level C, $15.75 for Level D and Level E, $17.15 for Level F, and $17.85 for Level G. Levels D-G have separate answer sheets which come in groups of 50 and cost $5.00 for Level D, $7.50 for Levels E, F, and G. The Diagnostic Mathematics Inventory Interim Evaluation Tests come in six levels and a kit of testing materials for 32 students is sold at each level. The Level A kit costs $55.00, Level B kit costs $60.00, Level C kit costs $75.00, Level D kit costs $65.00, Level E kit costs $70.00, and Level F kit costs $85.00. A multilevel specimen set is available for $15.00, a one-level specimen set costs $5.00. Scoring service costs $.90 and over for Levels A-C, $.60 and over for Levels D-G.

4. Date of Publication: 1975

B. Purpose and Nature of the Test:

The test is a criterion-referenced instrument for use with children in grades 1-8 and is designed as a multifaceted resource package to give specific information about individual students' strengths and weaknesses in 11 math skills categories. The DMI has from 37 to 179 multiple-choice items per level, with each item testing a separate objective.

C. Development of the Test:

Items and objectives for the test were selected based on the test author's analyses of math curricula currently being taught in a variety of basal math texts. A teacher's guide provides a rationale and history for the Diagnostic Mathematics Inventory. Specific information about how items were selected and how the test was developed is not available. Field testing information in terms of item difficulties, KR-20, point biserials, and test-retest reliabilities are available.

D. Administration and Scoring:

The test may be group administered by the classroom teacher. Test items are all multiple choice. The inventory is machine scored. The interim tests must be hand scored. Time to administer the test ranges from 80 minutes for Levels A-C, to 195-295 minutes for Levels D-G. Time required to administer the Interim Tests ranges from 10-25 minutes per test.

E. Analysis and Interpretation of Results:

No information is available on the range of difficulty of items at any level. The Inventory itself consists of 325 behaviorally stated objectives and test results indicate which have been or have not been mastered. A teacher's guide is available which has a cross-reference inventory. Learning Activity Guides are also available which include lists of premastery behaviors and suggestions for classroom activities to overcome errors. A guide to non-textbook materials cross-refers objectives and inventory items with various materials, activity cards, and other published materials. A large number of the 325 objectives deal with arithmetic computation, memory of procedures, nomenclature and associations.

F. Evaluation:

The instrument is essentially a collection of items in inventory form closely tied to the kinds of skills that are typically taught in basal math texts. The test would have to be locally evaluated to determine how closely its content matches the curricular needs of a particular location.

Diagnosis: An Instructional Aid--Mathematics

A. General Information

1. Author(s): J. Guzaities, J.A. Carlin, and S. Juda

2. Publisher: Science Research Associates, Inc.
155 Wacker Drive
Chicago, Illinois 60606

3. Cost: Level A, including the survey test, costs $60.00 per set of materials for 30 pupils. Level B, including the survey test, costs $65.50 per set of test materials for 30 pupils. The survey test costs $6.30 per 30 tests, and $7.00 for 30 separate answer sheets.

4. Date of Publication: 1972-1973

B. Purpose and Nature of the Test:

The test is to be used with students in grades 1-6 to determine whether or not certain mathematical skills have been mastered and to direct the student to a number of texts and learning kits to study the relevant material. The two levels cover 581 specific-objective subtests in these five areas: computation, geometry, measurement, operations and problem solving, sets and numerations. At each level there is a survey test and a series of diagnostic probe tests. All items are multiple-choice.

C. Development of the Test:

No information on standardization, item selection, construction and test development is provided.

D. Administration and Scoring:

Little information on how to administer and score the tests is available. Presumably, tests may be group administered and pupils, at least at level B, are to read directions on how to complete test items. A survey test may initially be administered prior to the diagnostic test. Tests within a given level may be administered in any order.

E. Analysis and Interpretation of Results:

No scores are obtained from the tests or subtests. If students get both items on the survey test correct, no further work is considered necessary. If 1 of 2 items is missed, the teacher is to "counsel" with the student to determine whether to give the diagnostic probe in that specific objective area. If the student misses both items on the survey test, he should "probably" be given the diagnostic test in that area. The teacher must decide from missed items which specific objectives need further study. Each specific objective is related through the prescription guide to appropriate work in several texts and learning kits.

F. Evaluation:

No data on reliability or validity are provided. Individual users would have to evaluate the relevance, balance, and appropriateness of the objectives based on their individual curricular program needs. The prescription guides could be helpful to the teacher in planning programs, provided access to a variety of texts and materials exists.

Diagnosis: An Instructional Aid--Reading

A. General Information

1. Author(s): A.N. Shub, J.A. Carlin, R.L.Friedman, J. Kaplan, J. Katien

2. Publisher: Science Research Associates, Inc. (SRA)
155 North Wacker Drive
Chicago, Illinois 60606

3. Cost: The cost varies according to level selected. Reading level A (Grades 1-4) consisting of test materials for 25 students, including a survey test, costs $119.50. The survey tests alone cost $10.80 for 25 tests. Reading level B (Grades 3-6), consisting of test materials for 25 students, and a survey test, costs $87.50.

4. Date of Publication: 1974

B. Purpose and Nature of the Test:

The test is to be used to diagnose reading skills of pupils in grades 1-6, and to suggest prescriptions for the remediation of identified weaknesses. The following skill areas are included: phonetic analysis, structural analysis, comprehension, vocabulary, study skills, and use of sources. Each level has a survey test and a series of diagnostic probe tests. Item formats include multiple-choice, matching, fill-in, and ordering.

C. Development of the Test:

No information on standardization, item selection, construction or test development is provided.

D. Administration and Scoring:

The test may be group-administered by the classroom teacher. A survey test may be initially administered to determine if a diagnostic test in that skill area should be given. Portions of the survey test may be administered based on teacher judgment. Each diagnostic test is a single 4-page booklet on which pupils work directly on the front and back. A carbon interleaf reproduces the answers on the key on the inside. This does not allow the test taker to erase. Directions for scoring are clearly specified in the handbook. Parts of the test are given orally. Time required to administer the tests was not reported.

E. Analysis and Interpretation of Results:

The survey test consists of usually one or two items in each skill area. A student getting one of the two items correct should be checked to see if a diagnostic test should be administered. If both items are missed, it's stated that the diagnostic test should "probably" be given. Directions for using diagnostic results suggest only that a pupil may be weak in a skill in which one or more items are missed. Other information on result interpretation is not provided.

F. Evaluation:

Information on reliability, validity or parallel form comparability is not available. It is difficult to determine whether objectives being tested are consistent with those to be learned. Since so little basic data is available, it seems doubtful that this test could be useful to classroom teachers at this point in the test development.

Fountain Valley Teacher Support System in Reading

A. General Information

1. Authors: Richard L. Zweig and Associates

2. Publisher: Richard L. Zweig Associates
Testing Division
20800 Beach Blvd.
P.O. Box 73
Huntington Beach, California 92648

3. Cost: Fifty sets of tests for any one grade level cost $25.75-$43.25 for hand scored, and $49.50-$104.50 for self scored tests. A set of scoring stencils costs $9.00-$19.00 Fifty record forms cost $6.50. A set of cassettes and manual for any one level cost $51.25-$99.50. An administration manual costs $15.00.

4. Date of Publication: 1975

B. Purpose and Nature of the Test:

The tests are to be used with children in grades K-6 to assess reading skills in 5 curricular areas: comprehension, phonetic analysis, structural analysis, study skills and vocabulary development. Scores are to be used to diagnose, prescribe and judge mastery of the various reading skills tested. The number of objectives varies from 125 at level K-1 to 33 at level 4. There are from 2 to 12 multiple choice items per objectives, with an average of about 3 items.

C. Development of The Test:

Field testing was done with 10,000 students, but no results were reported.

D. Administration and Scoring:

Tests are group administered by use of a tape cassette with instructions for taking the test on tape. Each student is to record his/her answers on either the self scoring answer sheets or hand scoring sheets. Children may be placed at a level to begin testing either based on a grade equivalent score obtained on a standardized test or based on teacher judgement of past performance. Hand scoring tests may be scored using a key overlay. Self scoring tests are scored by removal of the special backing by the teacher in order to reveal the correctness of student responses. Time to take the test ranges from 5-25 minutes per test.

E. Analysis and Interpretation of Results:

Mastery is defined as 100% for 2 and 3 item tests and 67% -88% for subtests of other lengths. If it is decided that a student needs further study on particular objectives, "prescriptions" can be developed from consulting the list of teaching materials that have been keyed to each objective. This manual is called the Teaching Alternatives Supplement. There is no supporting data or information for grade placement of specific objectives or for the mastery standards that were set.

F. Evaluation:

No reliability or validity data is supplied. Many of the objectives included seem to be linked to the lower two levels of the program, and seem to focus on the more easily measured objectives. Thus the test fails to reflect a complete approach to reading assessment. This instrument needs more technical data on development before it is of much use to classroom teachers.

Fountain Valley Teacher Support System in Mathematics

A. General Information

1. Authors: Richard L. Zweig and Associates

2. Publisher: Richard L. Zweig Associates
Testing Division
20800 Beach Blvd.
P.O. Box 73
Hunting Beach, California 92648

3. Cost: The cost for 50 sets of tests for any one grade level ranges from $16.50-$46.50 for hand scoring tests, $60.50-$178.50 for self-scoring tests. Sets of scoring stencils cost $11.00-$31.00. Fifty record forms cost $6.50. A manual, prescription guide and set of cassettes for any one level cost $83.50-$203.00.

4. Date of Publication: 1972

B. Purpose and Nature of the Test:

This criterion-referenced instrument is to be used with children in grades K-8 to assess 786 specific-objectives in 9 areas of math. The number of objectives per level ranges from 36 at the lowest to 135 at grade 6. The number of multiple choice items per objective ranges from 2 to 12.

C. Development of the Test:

The test materials include no information on how test items were constructed, selected or developed. No field testing information is available.

D. Administration and Scoring:

The test may be group administered by use of a tape cassette, which presents the test instructions to examinees. The teacher's manual includes directions for administration and scoring and for completing student profiles. Test items are either multiple choice or open-ended. Tests may be either self-scored, where the teacher removes a special back from the self scoring form to reveal the correctness of the student response, or hand scored with an overlay key. 11-25 minutes are required to administer each test.

E. Analysis and Interpretation of Results:

A mastery level is set at 67% for 3 item subtests and at 75% for 4 item subtests, with a range of 57% to 100% for subtests of other lengths. No supporting data for grade placement of specific objectives or mastery levels is included. Reteaching of objectives not passed is suggested

and an accompanying supplement contains references to materials published that pertain to specific objectives.

F. Evaluation:

No validity or reliability data is available. The cost of the test seems quite high, particularly in the absence of supporting data to substantiate its usefulness and technical values. This test does not seem appropriate for consumer use at this time.

Individualized Criterion-Referenced Testing: Math

A. General Information

1. Authors: Publisher

2. Publisher: Educational Progress
Educational Development Corporation
P.O. Box 45663
Tulsa, Oklahoma 74145

3. Cost: The total cost is $27.60 for 10 sets of all tests for any one level, 50 answer sheets, one scoring stencil, 10 individual record folders and one manual. A scoring service from the publishing company is also available at a cost of $1.35 per student.

4. Date of Publication: 1973

B. Purpose and Nature of the Test:

The test is designed to be used with children in grades 1-8. The test consists of 8 levels including 39 tests covering 312 overlapping specific-objective subtests of 2 items each in the math subject area.

C. Development of the Test:

Field testing reportedly was done in California, but no data was reported.

D. Administration and Scoring:

The test can be administered by the classroom teacher to the group. Students in grades 1 and 2 may use separate answer sheets. Students in grades 3 and up must use separate answer sheets. The tests can either be scored by hand or submitted to NCS scoring service.

E. Analysis and Interpretation of Results:

The systems approach of the test's authors included pretesting; scoring, reporting and prescribing instructional materials; instruction; posttesting. Only 2 items are included in each subtest. Mastery is defined as answering both items correctly. Need to review is defined as missing 1 of 2 items. When both items are missed, the student is identified as needing to learn the material corresponding to that specific subtest.

Following the publishers scoring of the pretests, the school district receives a report for each class including each objective tested, and the names of students who need to learn or review the objective. Individual student summaries are also provided showing which objectives

have been mastered, which need to be reviewed and which need to be learned.

F. Evaluation:

No validity or reliability was reported. It is questionable whether the classroom teacher could select appropriate booklets containing appropriate objectives for the particular class being taught. Many of the items tested are also no longer being taught in the present day curriculum. With the test being constructed with only 2 multiple choice items representing an entire subcategory of math, it is also questionable whether the test reliably reflects pupil achievement and knowledge.

Individualized Criterion-Referenced TEsting: Reading

A. General Information

1. Author(s): Publisher

2. Publisher: Educational Progress
Educational Development Corporation
P.O. Box 45663
Tulsa, Oklahoma 74145

3. Cost: The total cost is $27.60 for 10 sets of all tests for any one level, 50 answer sheets, one scoring stencil, 10 individual record folders and one manual. A scoring service from the publishing company is also available at a cost of $1.35 per student.

4. Date of Publication: 1973

B. Purpose and Nature of the Test:

The test is designed for use with children in grades K-8, and consists of 9 separate levels. Reading skills measured include word attack, literal comprehension, and interpretative comprehension. The manual suggests that the test be used for diagnosis and prescription, but no specifics as to how to use the test scores in this manner are provided. The manual suggests an alternative use of the test as a pretest and a posttest.

C. Development of the Test:

The items of the original test form were field tested on 80,000 students in grades 1-8 in Orange County, California. The test items were then revised. No further field test information was provided on the revised items. No information is provided on how the items were selected and sequenced.

D. Administration and Scoring:

The tests can be group administered by the classroom teacher. No information on which levels should be administered to which pupils was provided. Information on time required for administration was not available. The test can be both hand- and machine-scored.

E. Analysis and Interpretation of Results:

An operational skills survey was to be administered to students to determine the child's readiness to enter the testing program. No further information was provided to direct the teacher as to which level to administer to students. Information on the relationship of the objectives and the test items to the reading process was unavailable. Score reports provided good information to the teacher on each child's performance on the test items.

F. Evaluation:

The manual discussed definitions of validity and reliability. However, information concerning this test's validity and reliability was not specifically provided. More information on the relationship of test performance to reading skills needs to be available before the test can be considered useful for the classroom teacher.

Individual Pupil Monitoring System-Mathematics

A. General Information

1. Author(s): Riverside Publishing Co.

2. Publishers: Houghton Mifflin Company
Test Department
P.O. Box 1970
Iowa City, Iowa 52240

3. Cost: $15.45 for 35 tests and pupil progress records for any one test of Level 1, $18.15 for 35 tests and pupil progress records for any one test of Level 2-8, $21.00 per 500 hand scored answer sheets, $11.70 per 100 instamark answer sheets, $9.90 for Level 1 teacher's kit, $4.95 for Level 2-8 teacher's kits. Specimen sets are available for $3.90 per set.

4. Date of publication: 1973-1974

B. Purpose and Nature of the Test:

The test is designed as a criterion-referenced test for use with children in grades 1-8. It is a comprehensive set of instruction-referenced tests with each test focusing on a specific objective. The entire system provides separate tests for a total of 442 objectives. The number of objectives ranges from 48 at level 1 to 64 at level 8, the lower 3 levels having 5 multiple choice items per objective and the upper levels having 10. The test is designed to monitor the achievement status of pupils and summarize their performances in math which they have been rehearsing in day-to-day instruction.

C. Development of the Test:

Items of the test were taken directly from existing published mathematics material. A reference list is provided with the program--all programs referenced were published prior to 1974. The organizational format of the test predominantly follows the order of presentation in the 1967-70 edition or 1972 edition of the Modern School Mathematics series published by Houghton Mifflin. Field testing reportedly was done with a national sample of about 350 students for each level.

D. Administration and Scoring:

The IPMS tests are grouped into 8 skill areas: numeration and number systems, basic mathematical operations, geometry, measurement, problem solving, probability and statistics, and sets. At each level, the objectives and tests are divided into 3 parts which supposedly follow the order in which the skills are ordinarily taught during the school year. The parts are to be administered separately -- part A in the fall, part B in the winter and part C in the spring. The tests can be group administered by the classroom teacher. The tests may be hand scored or, at levels 3-8, self scoring answer sheets requiring the use of special crayons may be used. Time required for administration ranges from 60-150 minutes.

E. Analysis and Interpretation of Results:

The teacher's kit consists of a teacher's guide, a teacher's objective management record for each level, a reference booklet and behavioral objectives. No criteria for mastery of each objective are provided.

F. Evaluation:

No information on reliability is provided. Since the items on the tests look a lot like problems in existing mathematics programs, the test appears to have face validity. However, because all referenced programs were published prior to 1974, and current programs have changed some in structure, the test may need to be updated. If the instructional program being used by a school system is the Modern School Mathematics series, the IPMS could be useful for monitoring pupil progress. However, district using other materials need to carefully evaluate the actual match of specific test matter with their specific instructional program prior to selecting this system for use.

Individual Pupil Monitoring System--Reading

A. General Information

1. Authors: Riverside Publishing Co.

2. Publisher: Houghton Mifflin Company
Test Department
P.O. Box 1970
Iowa City, Iowa 52240

3. Cost: $18.50-$19.50 per 35 tests and pupil progress records, $21.00 per 500 hand scored answer sheets, $11.70 per 100 insta-mark answer sheets, $3.90 per box of 12 insta-mark crayon, $3.90 per teacher's kit for Level 1 or 2, $3.75 for any level teacher's kit for Levels 3-6, $1.80-$2.70 for cross-reference booklet. A specimen set is available for $3.90.

4. Date of Publication: 1974

B. Purpose and Nature of the Test:

The system is a criterion-referenced test consisting of 343 overlapping subskills in reading divided into 3 major areas -- word attack, vocabulary and comprehension and discrimination/study skills. The program is to be used with students in grades 1-6. There are from 43 to 63 objectives per level, with 5 multiple-choice items each level.

C. Development of the Test:

According to the test writers, prevalent behavioral objectives in use in reading were selected and arranged in 6 levels of difficulty. Qualified people were selected to write test items. Experimental editions were assembled and test tryouts were conducted on a national sample of students. Revisions were then made for the final test forms. No specifics on sample characteristics, on qualified reading experts or on how reading items were selected were given. No technical information was provided about field testing.

D. Administration and Scoring:

The test is to be administered to a group by the classroom teacher. Separate test booklets are provided for each reading area at each grade level. Answers may either be hand scored or self-corrected using the special insta-mark answer sheets and special crayons. Time for administration ranged from 90-165 minutes in area 1, 90-180 minutes for area 2, and 60-150 minutes for area 3.

E. Analysis and Interpretation of Results:

No suggested standards of mastery are stated and the manual also suggests that teachers should decide what level of mastery to use based on individual pupil skills.

F. Evaluation:

No data on test reliability is provided. Many of the test items are of questionable validity. It is doubtful whether the tests that are provided to measure attainment of some of the objectives are valid indices of the skills they purport to measure. Some of this lack of validity is due to improper format choice or improper test labeling. Because of its weakness in defining mastery and in relating some of the skills tested to real ability to read, the test would not be very useful to the teacher.

Key Math Diagnostic Arithmetic Test

A. General Information

1. Author(s): A. J. Connolly, W. Nachtman, and E. M. Pritchett

2. Publisher: American Guidance Service, Inc.
Publishers' Building
Circle Pines, Minnesota 55014

3. Cost: The cost for the complete kit is $26.50. The kit includes an Easel kit, a test manual, and 25 diagnostic records. A metric supplement test and manual are available for $4.25, and 25 metric supplement response forms are available for $2.50.

4. Date of Publication: 1971

B. Purpose and Nature of the Test:

The battery is designed for use with children in grades K-6 to diagnose deficit skills in math. At the level of specific objectives, there are 209 objectives, each with one test item. The items are grouped into 14 subtests which are in turn organized into three areas of math skills: content, operations, and applications. The number of items per subtest ranges from 7 to 27. Within subtests, items are grouped into "instructional clusters" of an average of 2 to 3 items.

C. Development of the Test:

Information on how test items were constructed and selected is provided in the manual. Behavioral objectives are also included. The test was constructed using the Rasch-Wright model. Initial field testing was done on a sample of about 2,000 children. The test was standardized on 1,222 children in K-7. No information is included regarding SES, parent education, and occupation levels.

D. Administration and Scoring:

The test is individually administered by using an Easel kit folder. It may be successfully used by classroom teachers or paraprofessionals after 5 or 6 practice trials. The problems mostly call for verbal answers to open-ended items presented orally and in combination with pictured illustrations. The test is scored as the subject takes each portion. It requires approximately 30-45 minutes to administer.

E. Analysis and Interpretation of Test Results:

The child's performance is evaluated on four levels: grade equivalent scores, scores related to content, operations and applications, profile showing the relative strengths in the 14 subtests, and analysis of performance on individual items. Deficit areas are delineated in sufficient detail to enable the teacher to write precise remedial prescriptions.

F. Evaluation:

The test has good reliability, total test reliability is .96, and good content validity, which ranges from .64 to .84. It was originally developed for use with educable mentally retarded children. It is quite complete and detailed. It could be very useful for planning of specific math programs with children in special education where individualized plans could be implemented based on the diagnostic information provided by the test.

Mastery: An Evaluation Tool-Mathematics

A. General Information:

1. Authors: SRA staff

2. Publisher: Science Research Associates, Inc.
155 North Wacker Drive
Chicago, Illinois 60606

3. Cost: $17.69 for 25 NCS scorable tests for grades K-2. $12.29 for 25 tests for grades 3-9, $10.71 for 100 answer sheets, 70¢ for each user's guide. A specimen set costing $10.00 is available. A scoring service costs $1.40 of less per test for grades K-2, 98¢ or less per test for grades 3-9.

4. Date of Publication: 1975

B. Purpose and Nature of the test:

The battery of tests is to be used with children in grades K-9 to evaluate skills at the recall and knowledge levels in 10 areas of math. The areas tested depend upon which age level is being tested and which areas are deemed appropriate by school personnel. There are 15 to 40 objectives per level with 3 multiple choice items per objective.

C. Development of the Test:

The guide describes the procedure used in developing the tests. Item difficulties, point biserials for each item, and KR-20 for each level are reported.

D. Administration and Scoring:

The manuals include straightforward, clear directions to the teacher for adminstration and to children for taking the tests. Instructions are read aloud by the teacher to the group. All answer sheets are machine scorable. Approximately 3 minutes per subtest are required for administration. Each test consists of 15-40 three-item single objective subtests.

E. Analysis and Interpretation of Results:

Perfect subtest scores indicate mastery of an objective. Customized tests covering locally chosen objectives are available. The manual includes a rationale for establishing mastery level at 3 out of 3 items.

F. Evaluation:

Content validity was established by reviewers, but procedures used to establish the validity are not described. Difficulty level of various items within the subtests is somewhat uneven. There is a significant jump in difficulty level from grade 2 to grade 3. Little problem solving or applicability is included in the test items content. The tests also do

not follow some of the trends in newer curricular development. The customized forms of this test battery could be used successfully by selecting those objectives which closely match local curricular content.

Mastery: An Evaluation Tool - Reading (SOBAR)

A. General Information:

1. Authors: SRA ataff

2. Publisher: Science Research Associates, Inc.
155 North Wacker Drive
Chicago, Illinois 60606

3. Cost: $17.69 per set of 25 scorable tests for grades K-2, $12.29 per set of 25 tests for grades 3-9. $10.71 for 100 answer sheets, 70¢ per user's guide. A specimen set is available for $8.20. A scoring service costs $1.40 or less per test at levels K-2, 98¢ or less per test at levels 3-9.

4. Date of Publication: 1975

B. Purpose and Nature of the Test:

SOBAR (System for Objective Based Assessment of Reading) is a criterion-referenced instrument to be used with children in grades K-9 for purposes of assessing mastery of a large number of reading skill behaviors in 6 general areas: letter recognition, phonic analysis, structural analysis, vocabulary, comprehension, and study skills. There are three multiple choice items per objective, the number of objectives ranging from 23 at level K to 35 at the upper levels.

C. Development of the Test:

No information is provided on test item selection, construction or item relation to reading performance. Item difficulty, point-biserials and KR-20s information is provided in a technical report.

D. Administration and Scoring:

The teacher is to use the manual, manual supplement and guide to determine how to administer the tests. In some cases, the teacher must switch back and forth from the manual to the supplement to the answer sheet in order to give specific portions of each test. Tests may be group administered. Each subtest takes approximately 3 minutes to give. The test consists of 23-35 subtest scores in 6 areas. Customized tests covering locally chosen objectives may be given or one of two forms of a catalog test of objectives may be chosen instead. Scoring services for the pupil, the class, or for a system provide information on each objective tested.

E. Analysis and Interpretation of Results:

Criteria for mastery - a perfect score on a 3 item subtest are arbitrarily set. No instructional suggestions are included other than a recommendation to use the probes, which include instructional prescriptions

comprised of suggested pages from basal textbooks and supplementary materials.

F. Evaluation:

Information on validity are not provided. A rationale for sampling of reading skill behaviors is also not provided. This test seems of questionable utility to the classroom teacher, who could probably make much better use of assessment materials accompanying the local curricular textbook series.

Prescriptive Reading Inventory

A. General Information

1. Author(s): CTB Staff

2. Publisher: CTB/McGraw-Hill
Del Monte Research Park
Monterey, California 93940

3. Cost: 35 CompuScan tests cost $20.65, 35 hand-scorable tests cost $16.45 for levels 1, 2, A, B, C. Level D costs $16.60 for 35 tests and 50 answer sheets when machine scored, $21.85 when hand scored. The PRI interim tests cost $75.00 for 32 sets of tests for Levels 1, 2, A or B, $80.00 per 32 sets of tests for Levels C and D. A multilevel specimen set is available for $10.00, a one-level specimen set costs $5.00. The CompuScan scoring service costs $.90 per booklet, $.60 per answer sheet with a $50.00 minimum order. $8.95 is the cost for each tape cassette for Levels 1 and 2.

4. Date of Publication: 1972

B. Purpose and Nature of the Test:

The Prescriptive Reading Inventory is for use with children in grades K-6 to assess individual students' achievement on a set of specific reading objectives in seven different reading areas. The Prescriptive Reading Inventory Interim Tests are to be used as a monitoring system following instruction based on the Prescriptive Reading Inventory. The Prescriptive Reading Inventory was designed to "provide evaluation relevant to classroom instruction." There are 10 objectives each for Levels 1 and 2 and 34 to 42 objectives per level for the upper four levels. The Interpretive Handbook includes guidelines for developing instructional objectives and activities based on PRI results. There are 3 to 5 test items for each objective.

C. Development of the Test:

90 objectives for the test were selected from a review of five basal reading programs. No theory of reading was used as a guide for the selection of objectives. The objectives are organized into major categories for the purpose of making the objectives easier to work with. Many of the objectives overlap. All five of the basal readers on which the PRI is based have been replaced with newer editions since this test was developed. Field testing was done with 18,000 students. Reliability and validity data were provided.

D. Administration and Scoring:

The test may be group administered by the teacher of reading. Levels 1 and 2 may be administered by tape cassette. The tests may be hand or machine scored. Time required to administer the Prescriptive Reading Inventory ranges from 75 to 200 minutes. Time required to administer the Prescriptive Reading Inventory Interim Tests ranges from 5 to 10 minutes per skills test, 20 to 25 minutes per comprehension booklet. The test can be scored by the scoring service or by hand scoring keys provided in the Interpretive Handbook for all levels.

E. Analysis and Interpretation of Results:

Three types of results are described--mastery, needs review, and non-mastery. Mastery on the PRI is defined as 66-2/3 percent and 75 percent correct on 3 and 4 items, respectively. Mastery on the PRI Interim Tests is defined as 100 percent correct for 3 item tests and 80 percent for 5 item tests. The unifying characteristics among the skills tested are not made evident by the test authors. No data are available to support the use of the scores purported to indicate mastery of particular objectives, nor is there sufficient data to support the ability of the tests to provide enough information for grouping of students for instructional purposes.

F. Evaluation:

Due to the limited number of test items per objective, reliability seems doubtful. Insufficient information is currently available for consumers to judge the concurrent and diagnostic validity of the tests. It would be necessary for the test user to carefully inspect the items and test format for each of the 90 objectives used. This time might be better spent using informal assessment as part of the ongoing instructional program which is currently in use within his or her particular classroom.

Woodcock Reading Mastery Tests

A. General Information

1. Author: Woodcock, R. W.

2. Publisher: American Guidance Service (AGS)
Publisher's Building
Circle Pines, Minnesota 55014

3. Cost: The test is available in two forms, costing $22.00 per kit of either form, 25 response forms and manual.

4. Date of Publication: 1973

B. Purpose and Nature of the Test:

The tests are designed to precisely measure reading ability in children in grades K-12. Reading abilities measured include letter identification, word identification, word attack, word comprehension, and passage comprehension.

C. Development of the Test:

The George Rasch analysis procedures served as a model for the development of the test items. Items on both forms of the test were selected from pools of items which met both statistical tests in the Rasch analysis. A detailed description of test development is provided in the manual. Because of the "item-free" nature of the Rasch process, only a small representative sample of items at each difficulty level had to be used for norming purposes. The final samples for norming included a nationally representative group of children from K-12. Norms are clearly presented, thoroughly researched and well constructed.

D. Administration and Scoring:

The test is to be individually administered in a session of 30 - 50 minutes. Though the manual suggests that the test may be administered by paraprofessionals, much of the manual itself includes highly complicated and technical data which needs to be interpreted by a reading or assessment professional. Actual test administraiton involves use of a convenient easel format. All required test responses are open ended. The manual presents precise directions for scoring the tests, though the scoring of various tests and the calculation of grade equivalents mastery scores, etc. may appear a bit complicated with initial use of the test. It is necessary for the examiner to spend quite a bit of time in learning how to use the many norm-referenced and criterion-referenced score reporting choices presented. An extensive set of tables is provided for interpretive options.

E. Analysis and Interpretation of Results:

The test user is provided with traditional as well as innovative reporting procedures. Six scores are recorded plus derived scores in these six areas at each of four levels: easy reading level (96% mastery), reading grade score (90% mastery), failure reading level (75% mastery), and relative mastery of grade level (individual). Newer interpretation procedures include (1) Relative mastery - this suggests an instructional range, (2) the achievement index - this suggest the degree of reading retardation and (3) the relative mastery at grade - this predicts the degree of success a subject would have when given a test similar to those that an average pupil at the subject's grade could perform with 90% mastery. Though the test authors claim that the test is criterion-referenced as well as norm-referenced, there is little information provided to enable it to be used as a criterion-referenced instrument. Procedures for interpreting a subject's scores in terms of his socio-economical status is involved and requires the collection fo data for eleven factors. The adjustment procedure described may be so time consuming that it is unfeasible.

F. Evaluation:

Corrected split-half reliabilities for four of the subtests range from .83 to .99. Above the 2.9 grade level, the reliability of the Letter Identification test drops sharply and becomes negligible at the 7.9 level. The standard errors of measurement are generally quite acceptable. No validity studies involving external criteria are reported. The authors claim predictive validity for the test based on using alternate forms of the test in a test-retest procedure. This could more correctly be described as a procedure establishing test-retest reliability. Three of the five subtests are of a type often used in reading diagnosis. These tests could be a valuable diagnostic tool for the experienced reading diagnostician.

BEST COPY AVAILABLE

Wisconsin Design for Reading Skill Development: Word Attack

A. General Information

1. Author(s): Otto, W., Kamm, K., et al.

2. Publisher: NCS Educational Systems
4401 West 76th Street
Minneapolis, Minnesota 55435

3. Cost: Level A, grades K-2, costs $17.00 per 35 consumable booklets and manual, $16.00 per set of ditto masters. Level B, grades 1-3, costs $13.50 per 35 consumable booklets and manual, $11.00 per set of ditto masters. Level C, grades 2-4, costs $20.50 per 35 consumable booklets and manual, $17.00 per set of ditto masters. Level D, grades 3-6, costs $10.00 per set of 35 consumable booklets and manual, $7.00 per set of ditto masters.

4. Date of Publication: 1972

B. Purpose and Nature of the Test:

The test is designed for use with children in grades K-6 to aid in instructional planning and evaluation. Information from the test results is to be used for determining the level and emphasis in instruction. The tests are the assessment portion of the Wisconsin Design for Reading Skill Development and are to be used in conjunction with that material, as well as with selected published materials, referenced in the Wisconsin Design. The objectives deal mainly with readiness, phonics, sight reading, and structural analysis. It is a four-level battery, with 6 to 16 objectives per level and at least 15 multiple-choice items per objective.

C. Development of the Test:

Little specific information on how test items were constructed is available. The test authors constructed test items to measure word attack skills which they and classroom teachers had identified as important. Field testing was done with a median of 152 students per level. A variety of data are given, including, for each objective, average correct, frequency distributions, and internal consistencies.

D. Administration and Scoring:

Manuals contain clear and adequate instructions to teachers for administering the tests to groups. Recommendations for appropriate

sizes of groups being tested are included for each level. Teachers can select any particular sections to administer to individual students. The tests are untimed; suggestions for pacing particular tests are included. The estimated time for testing a single skill is 12 minutes. Four class periods are required for administering the first three levels, five periods for Level C, and two periods for Level D. Raw scores, percentage correct and mastery scores are derived. Mastery is to be determined in accord with the local criterion, however a suggested mastery level is 80 percent for any particular subtest given.

E. Analysis and Interpretation of Results:

This test is to be used in instructional planning. Eighty percent mastery criterion is suggested for each subtest. If a child fails not more than one subtest, he is to be retested at the next higher level. If the child passes not more than one subtest, he is to be retested at the next lower level. Six to sixteen single-skill scores can be derived at each of five levels. The tests are very comprehensive in terms of including almost every reading behavior which comprises the word attack skill area. Thus, test results are to be used to make a diagnostic evaluation of each child's strengths and weaknesses. Formal assessment following instruction in areas identified as weaknesses is recommended.

F. Evaluation:

The test manuals state that reliabilities of individual subtests are in the .70s and .80s with a few in the .90s. They state that further specific information appears in the technical manual (which was never published). Validity information is claimed on the basis of test items being tied to behavioral objectives identified by the authors as measuring word attack skills. Since the areas tested appear to be quite comprehensive in scope and the test materials offer much information on available instructional materials to teach skills tested, this test could be a valuable aid to the classroom teacher in focusing on needed areas of instruction in reading--word attack skills.

Wisconsin Tests of Reading Skill Development: Study Skills

A. General Information

1. Author(s): Otto, W., Kamm, K., et al.

2. Publisher: NCS Educational Systems
4401 West 76th Street
Minneapolis, Minnesota 55435

3. Cost: Level A, grades K-1, costs $10.00 for 35 consumable booklets and a manual, $6.00 for a set of ditto masters. Level B, for grades 1-2, costs $17.00 for 35 consumable booklets and a manual, $15.00 for a set of ditto masters. Level C, grades 2-3, costs $24.50 for 35 consumable booklets and a manual, $25.00 for a set of Form P ditto masters, $1.00 for a special Form Q ditto master, $15.00 for 35 reusable color maps booklets. Level D, for grades 3-4, costs $28.00 for 35 consumable booklets and a manual, $28.00 for a set of Form P ditto masters, $12.00 for a set of Form Q ditto masters. Levels E-G, grades 4-5, 5-6, 6-7, costs $60.00 for 35 tests, $3.00 for a set of ditto masters for answer sheets. A sampler is available free.

4. Date of Publication: 1973

B. Purpose and Nature of the Test:

The Wisconsin Tests of Reading Skill Development: Study Skills are designed for use in conjunction with the Wisconsin Design for Reading Skill Development. The test is a formal, criterion-referenced instrument composed of sevel levels for use with children in grades K-7. There are from 2 to 14 objectives per level, with at least 10 multiple-choice items per objective. The battery is to be used to determine a student's skill proficiencies and deficiencies prior to instruction, and to monitor progress after instruction.

C. Development of the Test:

According to the manual, skills selected for the test are based on input from teachers, research results and a "common sense analysis" of tasks of elementary school students. Norms are not available. The tests focus on eight "strands" which include the skills necessary for locating and reference materials. Field testing was done with more than 1000 students. A variety of data are given including, for each objective, average correct, frequency distribution, and internal consistency.

D. Administration and Scoring:

Manuals provide directions that appear to be clear and appropriate to the ages being tested. Tests may be group-administered by classroom teachers. Tests are untimed, and require from 25 to 190 minutes to administer with the shorter times corresponding to lower age levels. Examiners are encouraged to help individual children with further explanations as required.

E. Analysis and Interpretation of Results:

The tests are referenced to specific scored objectives. Scoring is percentage correct with 80 percent signifying mastery of a strand. Mastery of a higher level strand implies mastery at lower levels. If a child fails not more than one subtest, he is to be retested at the next higher level. If he passes not more than one subtest, he is to be retested at the next lower level. Since the tests closely correspond to the Wisconsin Design, areas of deficiency can be remedied by using instructional materials from the Wisconsin Design for Reading Skill Development.

F. Evaluation:

Validity and reliability information is somewhat sparse. The manual states that reliabilities of individual tests are in the .70s and .80s, with a few in the .90s. Sample sizes for reliability judgments were quite small and some test reliabilities were much lower than those emphasized. Seventeen tests had reliabilities below .70--one was as low as .32. The manual refers to a technical manual with more specific technical information, but no technical manual was available. The test objectives do correspond closely to material provided in the Wisconsin Design. From examining the Wisconsin Design Material objectives, teachers could determine if its content was relevant to their class curriculum. If the content covered is considered of importance, then these tests would provide valuable monitoring information on students' progress in this skill area.

Stanford Diagnostic Mathematics Test

A. General Information:

1. Author(s): L. S. Beatty, et al.

2. Publisher: The Psychological Corporation
Harcourt Brace Jovanovich
757 Third Avenue
New York, New York 10017

3. Cost: The cost varies according to the level of the test selected. Level a (grades 1.5-4.5) costs $12.50 per 35 tests, $3.50 per set of hand-scoring stencils. A scoring service may be used. The cost for MCR-scored tests is $13.50 per 35 tests, $.85 and over per test for MCR scoring service. Levels b-d (grades 3.5-6.5, 5.5-8.5, 7.5-13) cost $12.50 per 35 hand-scored tests, $2.95 per 35 hand-scored answer sheets, $3.25 per 35 MCR answer sheets, $3.50 per set of hand-scoring stencils for test booklets, $1.25 per hand-scoring stencil for answer sheet. MCR scoring service is $.80 and over per test. No prices for manuals were given. A specimen set is available for $3.00.

4. Date of Publication: 1976

B. Purpose and Nature of the Test:

The test is designed to identify strengths and weaknesses of the individual pupils in three areas of mathematical competence: concepts of the number system and numeration, skill in computation, and application of these concepts and skills to problem-solving situations. The test is to be used with children in grades 1.5 to 13. There are four levels in the test, with 11 to 13 objectives at each level. The number of test items per objective range about 8-10. The test can be used for identifying groupings for instructional purposes.

C. Development of the Test:

National norms for the test were established and seem to have been appropriately and adequately accomplished. More specific information on standardization is available from the test manual. The test authors' judgment on item selection and mastery cutoff scores was "guided by the relative importance of the measured concepts and skills and their location in the instructional sequence relative to pupils' grade placement and by the performance of pupils at different achievement levels."

D. Administration and Scoring:

The test may be administered to the class by the classroom teacher. Separate answer sheets are to be used and may either be hand-scored or sent to the MCR scoring service. At level a, 2 of the tests are read aloud to the students, and the questions are read only once. This could cause problems for poor auditory learners, since the text being read aloud is not included in testing booklets for level a. Time for administering the test ranges from 95-120 minutes.

E. Analysis and Interpretation of Test Results:

The test is to be used to identify strengths and weaknesses in pupils' understanding of basic ideas of numeration and place value and skills in computation and applications needed before study of additional ideas in mathematics. "Progress indicators," a plus or minus indicating scores at or below a designated mastery cutoff score, are provided for item clusters. Subtest scores and item cluster scores can be obtained. The manual gives adequate explanations of the scores and includes a section on the implications of the test results for instructional purposes. Performance objectives are listed for each test.

F. Evaluation:

The internal consistency reliability and comparison correlations with the Stanford Mathematics Test seem acceptably high. The test, therefore, seems to be a valid and reliable instrument for helping classroom teachers identify general areas of strength and weakness among individual pupils in the subject area of mathematics. The test includes many easy questions to assure more precise measurement of below-average pupils. The test will provide a good overview of student skills and indicate a starting point for more in-depth evaluation and planning for individual math learning programs.

Stanford Diagnostic Reading Test

A. General Information:

1. Author(s): B. Karlsen, R. Madden, and E. F. Gardner

2. Publisher: The Psychological Corporation
Harcourt Brace Jovanovich
757 Third Avenue
New York, New York 10017

3. Cost: Red and green levels cost is $12.50 per 35 hand-scored tests, $14.95 per 35 MCR-scored tests, $3.50 per set of hand-scoring stencils. An MCR scoring service is available at a cost of $.90 and over per test. Brown level cost is $12.50 per 35 tests, $7.70 per 35 MCR answer folders, $3.50 per 35 hand-scored answer folders, $2.75 per set of hand-scoring stencils. MCR scoring service is available at a cost of $.85 and over per test. Blue level cost is $14.50 per 35 tests, $8.00 per 35 answer booklets, $3.25 per set of scoring stencils. Scoring service is available at a cost of $.60 and over per test. A specimen set is available at any level for $3.00.

4. Date of Publication: 1976

B. Purpose and Nature of the Test:

The test is designed to diagnose reading strengths and weaknesses in students from grades 1.5 to 13 and particularly to provide accurate assessment of low-achieving students. The test has 4 different levels which are identified by color. The Red level is for grades 1-2, the Green level is for grades 3-4, the Brown level is for grades 5-8, and the Blue level for grades 9-12. The four levels overlap somewhat in the skills they measure. There are 17-25 objectives at each level with generally 6-8 multiple-choice test items per objective. A handbook referencing the tested skills to a variety of reading series is offered.

C. Development of the Test:

Test items were selected based on the identification of the instructional objectives common to most reading programs found in state and city curriculum materials and major reading series. Based on this information, the authors designed the test made up of a hierarchical set of component skills sequenced according of order of complexity. The test was standardized on approximately 31,000 students in 3,000 school districts nationwide in 1975.

D. Administration and Scoring:

The manual gives clear, concise and complete instructions for administering the test. The test can be administered by the classroom teacher in a group setting. The test can be either hand-scored or sent to the MCR scoring service. Separate answer sheets or answer folders are used at various levels by students to record their answers. Time to administer the test ranges from 113-165 minutes and the test may be administered in 1-5 sessions, depending on which level is being given.

E. Analysis and Interpretation of Test Results:

The test features two types of scores: content-referenced scores and norm-referenced scores. Norm-referenced scores include percentile ranks, stanines, grade equivalents, and scaled scores. Content-referenced scores include raw scores and progress indicators. The test authors provide suggestions for selecting the test performance information most useful in the local school. This information is supposed to help teachers plan groupings for reading instruction.

F. Evaluation:

The reliability coefficients for this test ranged from .79 to .98 with many coefficients for various subtests exceeding .90. This test provides useful information to classroom teachers to aid in individualizing reading instruction in order to meet a variety of pupil needs.

References

Ahmann, J.S. and Glock, M.D. Evaluating student progress - Principles of tests and measurements. Boston, MA; Allyn & Bacon, Inc 1981

Alpert, J.L. and Trachtman, G.M. School Psychological consultations in the eighties. Paper presented at the annual meeting of the National Association of School Psychologists New York City, 1979

Anastas; A Psychological testing (4th ed) Toronto, Canada; The Macmillan Co. 1976

Bailey, D.B. Jr. and Harbin, G.L. Nondiscriminatory evaluation. Exceptional Children, 1980, 46, 590-596.

Bardon, J.J. & Bennett, V. School Psychology Englewood Cliffs, New Jersey: Prentice-Hall, Inc. 1974

Bergquist, C.C. and Graham, D.L. Developing Models for special education Evaluation Review, 1980, 4, 307-321

Berk, R. Determination of optimal cutting scores in criterion-referenced measurement. Journal of Experimental Education, 1976, 45, 4-9

Berk, R.A. The application of structural facet theory to achievement test construction Educational Research Quarterly, 1978, 3. 62-72

Block, J. Criterion-referenced measurements: Potential. School Review, 1971, 69, 289-298.

Block, J. Standards and criteria: A response. Journal of Educational Measurement, 1978, 15, 291-295.

Buros, O.K (Ed) The eigth Mental Measurement yearbook. Edison, New Jersey; Gryphon press, 1978

Cohen, J. A coefficient for agreement of nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46

Ebel, R. Content standard test scores. Educational and Psychological Measurement, 1962, 22, 11-17

Ebel, R. Criterion-referenced measurements; Limitations. School Review, 1971, 69, 282-288.

Ebel, R.L. A case for minmal competency testing. Phi Delta Kappan, 1978, 59, 546-549

Glaser, R. Instructional Technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 519-521.

Glaser, R. & Nitko, A. Measurement in learning and instruction. In R.L. Thorndike (Ed.), Educational Measurement. Washington: American Council on Education, 1971

Glass, G. Standards and criteria. Journal of Educational Measurement, 1978, 15, 237-261.

Goh, D.S. Graduate training in school psychology. Journal of School Psychology, 1977, 15, 207-218

Goh, D.S., Fuller, G.B., and Teslow, C. The practice of psychological assessment among school psychologists. Professional Psychology, 1981, 12, 6, 696-706

Gray, W.M. A comparison of Piagetian Theory and Criterion-referenced Measurement. Review of Educational Research, 1978, 48, 223-249

Hambleton, R. On the use of cut-off scores with criterion-referenced tests in instructional settings. Journal of Educational Measurement, 1978, 15, 277-290.

Hambelton, R.K. Test score validity and standard-setting methods. In R.A. Berk, (Ed) Criterion-referenced measurement: The state of the art. Baltimore, Maryland: The Johns Hopkins University press, 1980.

Hambelton, R.K. and Eignor, D.R. Guidelines for evaluating criterion-referenced tests and test manuals. Journal of Educational Measurement, 1978, 15, 321-327.

Hambelton, R., & Novick, M. Toward an integration theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Hambleton, R.K., Powell, S., and Eignor, D.R. Issues and methods for standard setting. In R.K. Hambleton and D.R. Eignor, A practitioner's guide to criterion-referenced test development, validation, and test score usage (Laboratory of psychrometric and Evaluation Research Report No. 70, 2nd ed). Amherst: University of Massachusetts, School of Education, 1979.

Hambelton, R.K. Swaminathan, H., Algina, J. & Coulson, D.B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1, 1-47.

Harris, C.W., Alkin, M.C. and Pophan, W.J. (Eds) Problems in criterion-referenced measurement. CSE Mongraph series in Evaluation, No. 3 Los Angeles: center for the study of evaluation University of California, 1974.

Harris, M. & Stewart, P. Application of classical strategies to criterion-referenced test construction. Paper presented at the annual meeting of American Educational Research Association, New York, 1971

Hemphill J. & Westie, C.M. The measurement of group dimensions. Journal of Psychology, 1950, 29, 325-342.

Hively, W., Patterson, H. & Page, S. A "universe-defined" system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 5, 275-290.

Jaeger, R.M. A proposal for setting a standard on the North Carolina High School Competency Test. Paper presented at the spring meeting of the North Carolina Association for Research in Education, Chapel Hill, 1978

Linn, R.L. Two weak spots in the practice of criterion-referenced measurement. Educational Measurement, 1982, 1, 12-14.

Meskauskas, J.A. Evaluation models for criterion-referenced testing: Views regarding mastery and standard setting. Review of Educational Research, 1976, 46, 133-158.

Messick, S.A. The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 1975, 30, 955-966.

Millman, J. Passing scores and test length for domain-referenced measures. Review of Educational Research, 1973, 43, 205-216.

Popham, W.J. An approaching peril: closed-referenced tests. Phi Delta Kappan, 1974, 56, 614-615

Popham, W.J. Selecting objectives and generating test items for objectives based tests. In C.W. Harris, M.C. Alkin, and W.J. Popham (EDs). Problems in criterion-referenced measurement. CSE Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation. University of California, 1974.

Popham, W.J. Educational evaluation. Englewood Cliffs, New Jersey: Prentice-Hall, 1975

Popham, W.J. Normative data for criterion-referenced tests? Phi Delta Kappan, 1976, 58, 593-594.

Popham, W.J. Criterion-referenced Measurement. Englewood Cliffs, New Jersey: Prentice-Hall, Inc. 1978

Popham, W.J. Domain specification strategies. In R.A. Berk (Ed) Criterion-referenced measurement: The state of the art. Baltimore, Maryland: The Johns Hopkins University Press, 1980.

Popham, W.J. & Husek, T.R. Implications of criterions-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Rovinelli, R.J. & Hambleton, R.K. On the use of content specialists in the assessment of criterion-referenced test item validity. Dutch Journal of Educational Research, 1977, 2, 49-60.

Salvia, J. and Ysseldyke, J.E. Assessment in special and remedial education Boston: Houghton Mifflin, 1978

Sanders, J.R. & Murray, S.L. Alternatives for Achievement testing. Educational Technology, 1976, 3, 17-23

Shephard, L.A. Measurements consequences of selected standard-setting models. In M.A. Bunda and S.R. Sanders (Eds). Practice and problems in competence-based measurement. Washington, D.C.: National Council on Measurement in Education, 1979.

Shepard, L. Standard setting issues and methods. Applied psychological measurement 1980, 4, 447-467

Shoemaker, D. Improving criterion-referenced measurement. Journal of Special Education, 1972, 6.

Standards for educational and psychological tests and manuals Washington, D.C.: American Psychological Association, 1974

Swaminathan, H., Hambleton, R.K., and Algina, J. Reliability of criterion-referenced tests: a decision-theoretic formulation. Journal of Educational Measurement, 1974, 11, 263-267

Swezey, R.W. Individual performance assessment: An approach to criterion-referenced test development. Reston, Virginia: Reston publishing Company, Inc. 1981

Swezey, R.W. and Pearlstein, R.B. Guidebook for developing criterion-referenced tests. A report prepared for the U.S. Army Research Institute for the Behavioral and Social Science. Reston, Virginia: Applied Science Associates, August 1975

Thorndike, E.L. The nature, purposes and general methods of measurement of educational products. The Measurement of Educational Products, Seventeenth yearbook of the National Society for the Study of Education, Part II (Bloomington, IL: Public School Publishing company, 1918), 16-24

Walker, C.B. Standards for evaluating criterion-referenced tests Los Angeles: Center for the Study of Evaluation: University of California, 1977 (Unpublished manuscript)

Wendt R.N. Kindergarten entrance assessment: Is it worth the effort? Psychology in the Schools, 1978, 15, 56-62.

Zieky, M.J. & Livington, S.A. Manual for setting standards on the basic skills assessment tests. Princeton, New Jersey: Educational Testing Service, 1977

Figure captions

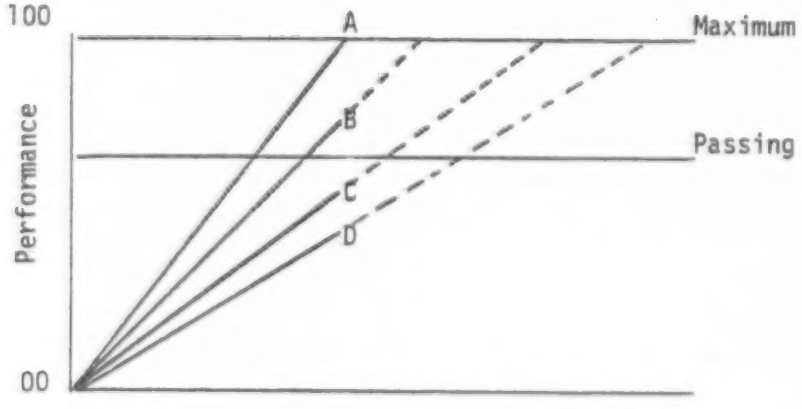
Figure 1. Performance levels of four different individuals (A, B, C, and D) on a criterion-referenced and a norm-referenced model.

Figure 2. Comparison of norm-referenced achievement tests with criterion-referenced tests.

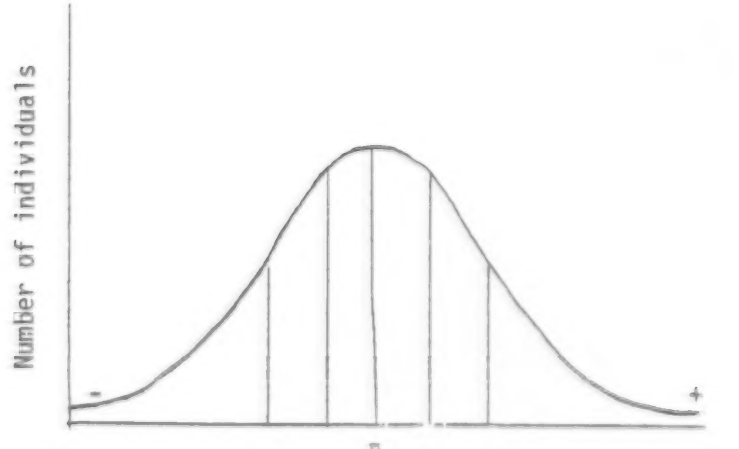
Figure 3. An example of domain and test specifications.

Figure 4. A classification of methods for setting standards.

Figure 1 Performance levels of four different individuals (A, B, C, D) on a criterion-referenced and a norm-referenced model



Criterion-Referenced Model



Norm-Referenced Model

Figure 2 Comparison of Criterion-Referenced and Norm-referenced Tests

Norm-referenced Tests	Criterion-Referenced Tests
1. Interpret test scores in relation to established norms.	1. Report which, or how many, of a set of specific achievement goals the individual has reached.
2. Sample the domain of a particular achievement area broadly.	2. Sample a limited number of specifically defined goals.
3. Provide a concise summary of less clearly defined areas of achievement.	3. Report specific and detailed information on pupil achievement.
4. Encourage and reward individual excellence in achievement.	4. Emphasize mastery of specific subject matter by all pupils.
5. Treat learning as consisting of building a structure of numerous relations between concepts.	5. Treat learning as if it were acquired by adding separate, discrete units to the collection of things learned.

Adapted from Ebel (1975).

Figure 3 An Example ofDomain and Test Specifications

An illustrative set of criterion-referenced test specifications: applying concepts of United States foreign policy

General description

Given a description of a fictitious international situation in which the United States may wish to act, and the name of an American foreign policy document or pronouncement, the students will select from a list of alternatives the course of action that would most likely follow from the given document or pronouncement.

Sample item

Directions: Read each fictitious example below. Decide what action the United States would most likely take based on the given foreign policy document. Write the letter of the action on your answer sheet.

Some Russian agents have become members of the Christian Democratic Party in Chile. The party attacked the president's house and arrested him. The Russian agents set themselves up as president and vice-president of Chile. Chile then asked to become an "affiliated republic" of the USSR.

Based on the Monroe Doctrine, what will the United States do?

a. Ignore the new status of Chile.
b. Warn Russia that its influence is to be withdrawn from Chile.
c. Refuse to recognize the new government of Chile because it came to illegally.
d. Send arms to all groups in the country that swear to oppose communism.

Stimulus attributes

1. The fictitious passage will consist of 500 words or less followed by the name of a foreign policy pronouncement or document inserted into the question, "Based on the _____, what will the United States do?
2. The policy named in the stimulus passage will be a document or pronouncement selected from the specification supplement.
3. Each passage will consist of two parts: (a) a background description of an action taken by a foreign nation and (b) a statement of the action to which the foreign policy document or pronouncement is to be applied.
    a. The background statement will be analogous to a historical situation that either preceded the issuance of the cited document or pronouncement or for which the document or pronouncement was used. For example, the Monroe Doctrine was drawn up in response to European designs on American nations that were attempting to establish independence. An analogous case today might describe a European country attempting to encroach on the sovereignty of an American country.
    b. The statement of an action will describe an action taken by a real foreign nation that conforms to one of the following categories:
        (1) Initiation of an international conflict
        (2) Initiation of a civil conflict. This may include coups, revolutions, riots, protest march-

es, civil war, or a parliamentary crisis.

(3) Initiation of an international relationship. This may include trade negotiations, friendship pacts, military alliances, and all classes of treaties.

(4) Appeal for foreign aid to meet economic or military needs.

(5) Development and stockpiling of military weapons.

4. All statements in the passage will refer to specific nations and events. Descriptions such as, "A nation is at war with another country," are not acceptable.
5. When the document or pronouncement mentioned in the stimulus passage is tied to a particular geographical region, countries named in the passage must belong to that region.
6. Passages will be written at no higher than the seventh-grade reading level.

Reponse attributes

1. Students will be asked to mark the letter of one of four given response alternatives consisting of the correct response and three distractors. Each alternative will possess the following characteristics:
    a. Describe a specific course of action that refers to the people, nations, and actions in the stimulus passage.
    b. Be brief phrases written to complete the understood subject, "The United States will . . ."
2. Distractors (wrong answers) will be written to meet these addi-

tional criteria:

a. Each distractor will describe an action derived from a different document or pronouncement selected from the specification supplement.

b. Documents or pronouncements from which identical courses of action may be derived will not be used.

c. The decision for the United States not to act may be used as a course of action when it is based on a document or pronouncement.

3. The correct response will be the course of action that is governed by the principles described in the document or pronouncement named in the stimulus passage.

Adapted from Popham (1978, pp. 129-131)

Figure 4 A Classification of Methods for Setting Standards

Judgmental Methods	Empirical Models	Combination Models
Item Content	Data-Criterion Measure	Judgmental-Empirical
Nedelsky (1954)	Livingston (1975)	Contrasting groups (Zieky & Livingston 1977)
Angoff (1971)	Livingston (1976)	Borderline groups (Zieky & Livingston 1977)
Modified Angoff (ETS, 1976)	Van der Linden and Mellenbergh (1977)	Criterion groups (Berk 1976)
Ebel (1979)		
Jaeger (1978)		
Guessing	Decision-Theoretic	Educational Consequences
Millman (1973)	Kriewall (1972)	Block (1972)
		Bayesian Methods
		Hambleton & Novick (1973)
		Schoon, Gullion, & Ferrara (1978)

Adapted from Hambleton (1980) in Berk (1980, p. 104).